People's Democratic Republic of Algeria

Algerian Ministry of Higher Education and Scientific Research

University of Frères Mentouri Constantine 1

Faculty of Science and Technology

Electronics Department

Submitted thesis, in partial fulfillment of the conditions
for the award of the 3rd cycle doctoral degree **LMD.**

Field: Automatic and signal processing

Presented by:
Oussama MESSAI

No-reference Stereoscopic Image Quality Assessment

Date : 16 / 12 / 2021

Presented in front of jury:

Pr. Noura MANSOURI	University of Frères Mentouri Constantine 1	**President**
Pr. Fella HACHOUF	University of Frères Mentouri Constantine 1	**Reporter**
Dr. Z. AHMED SEGHIR	University Abbès Laghrour Khenchela	**Co-Reporter**
Pr. Toufik BOUDEN	University of Jijel	**Examiner**
Pr. Salim CHIKHI	University Abdelhamid Mehri Constantine 2	**Examiner**
Pr. Atef FARROUKI	University of Frères Mentouri Constantine 1	**Examiner**
Dr. Aladine CHETOUANI	University of Orléans, PRISME laboratory	**Invited**

In the academic year: 2020/2021

Abstract

English :

Stereoscopic imaging is becoming increasingly popular, and its use in photography, television, and films is rapidly expanding. Obviously, access to this type of images often includes necessary treatments (acquisition, processing, compression, transmission, etc.), which may result in a variety of artifacts (blocking, blur, ringing, etc.). As a result, it is critical to have adequate tools for measuring the quality of stereoscopic contents.

It is thus essential to establish efficient metrics that assess the impact of these treatments on the perceived quality. To meet this critical need, significant efforts have been made to study and evaluate the quality of stereoscopic images. In this thesis, we present several contributions for quality assessment of stereoscopic contents. Five methods have been proposed in total, with all of them are no-reference based metric. These metrics were developed with Human Visual System (HVS) modeling and human visual attention (saliency information) in mind. In addition, various advanced techniques, such as deep learning, have been incorporated into our workflow designs.

Keywords: Stereoscopic Image Quality Assessment (SIQA), Human Visual System (HVS), Saliency information, Deep learning.

Français :

L'imagerie stéréoscopique est de plus en plus populaire et son utilisation dans la photographie, la télévision et les films se développent rapidement. Bien entendu, l'accès à ce type d'image comporte souvent des traitements nécessaires (acquisition, traitement, compression, transmission, etc.), qui peuvent se traduire par une variété d'artefacts (blocage, flou, bruit, etc.). En conséquence, il est essentiel de disposer d'outils adéquats pour mesurer la qualité des contenus stéréoscopiques.

Il est donc essentiel d'établir des métriques efficaces qui évaluent l'impact de ces traitements sur la qualité perçue. Pour répondre à ce besoin critique, des efforts importants ont été faits pour étudier et évaluer la qualité des images stéréoscopiques. Dans cette thèse, nous présentons plusieurs contributions pour l'évaluation de la qualité des contenus stéréoscopiques. Cinq méthodes ont été proposées au total, toutes étant des métriques sans référence. Ces mesures ont été développées en tenant compte de la modélisation du système visuel humain (HVS) et de l'attention visuelle humaine (informations de saillance). En outre, diverses techniques avancées, telles que l'apprentissage en profondeur, ont été intégrées à nos conceptions de flux de travail.

Mots clés : Évaluation de la qualité d'image stéréoscopique (SIQA), système visuel humain (HVS), informations de saillance, apprentissage en profondeur.

العربية :

أصبح التصوير المجسم ثلاثي الأبعاد شائعا بشكل متزايد في شتى المجالات ،كما أن استخدامه في التصوير الفوتوغرافي والتلفزيون والأفلام يتوسع بسرعة. من الواضح أن الوصول إلى هذا النوع من الصور غالبًا ما يتضمن المعالجات الضرورية (الاستحواذ والمعالجة والضغط والنقل وما إلى ذلك) ، والتي قد تؤدي إلى مجموعة متنوعة من تدهور الجودة (الحجب ، والتمويه ، والرنين ، وما إلى ذلك). نتيجة لذلك ، من المهم أن يكون لديك أدوات كافية لقياس الجودة في هذا النوع من الصور.

لذلك ، من الضروري وضع مقاييس فعالة لتقييم تأثير هذه العلاجات على الجودة. لتلبية هذه الحاجة الماسة ، تم بذل جهود معتبرة لدراسة وتقييم جودة الصور المجسمة. في هذه الرسالة ، نقدم العديد من المساهمات لتقييم جودة المحتويات المجسمة. تم اقتراح خمس طرق في المجمل ، وكلها لا تعتمد على المقاييس المرجعية. تم تطوير هذه المقاييس مع نمذجة النظام البصري البشري والاهتمام البصري البشري (معلومات البروز) في الاعتبار. بالإضافة إلى ذلك ، تم دمج العديد من التقنيات المتقدمة ، مثل التعلم العميق ، في تصميمات سير العمل لدينا.

الكلمات الدالة: تقييم جودة الصورة المجسمة ، النظام البصري البشري ، معلومات التميز ، التعلم العميق.

Acknowledgements

My deepest thanks, along with my entire gratitude, go to my promoter, HACHOUF Fella professor at the University of Constantine 1, who directed my thesis, generously guided me throughout this effort, and provided me with a framework for very beneficial work. I must express my deepest gratitude to my thesis's co-supervisor, Dr. ZIANOU Ahmed Seghir, for his support and remarks that aided me in taking my project ahead.

I particularly thank Dr. CHETOUANI Aladine for his inspiration and encouragement. I really appreciate his kind and professional feedbacks and discussions we had during my Ph.D internship at PRISME laboratory, University of Orléans.

A sincere thanks to Noura MANSOURI, professor at the University of Constantine 1, for having accepted to chair the jury and judge this work. I especially thank my PhD jury members for their interest in this work by agreeing to be examiners: Pr. Toufik BOUDEN, Pr. Salim CHIKHI, and Pr. Atef FARROUKI.

This work would not be possible without the support of my family, friends, and university employees from near and far. This thesis was carried out within the "Image et vision" team of the Automatic and Robotics laboratory (LARC), University of Constantine 1. I salute all of my comrades and colleagues from LARC and PRISME laboratory, I am so grateful for your support, concern and friendship, I wish you the best of luck in your future endeavors.

List of Abbreviations

HVS	Human Visual System
LGN	Lateral Geniculate Nucleus
CCD	Charge Coupled Device
CMOS	Complementary Metal Oxide Semiconductor
IQA	Image Quality Assessment
SIQA	Stereoscopic Image Quality Assessment metrics
JPEG	Joint Photographic Experts Group
WN	White Noise
FF	Fast Fading
SAD	Sum of Absolute Value
SSIM	Structural Similarity
AWGN	Additive White Gaussian Noise
ACR	Absolute Category Rating
DCR	Degradation Category Rating
MOS	Mean Opinion Scores
DMOS	Difference Mean Opinion Scores
VQEG	Video Quality Experts Group
RMSE	Root Mean Squared Error
PSNR	Peak-Signal-to-Noise-Ratio
PLCC	Pearson Linear Correlation Coefficient
SROCC	Spearman's Rank-Order Correlation Coefficient
KROCC	Kendall's Rank Order Correlation Coefficient
NSS	Natural Scene Statistics
QoE	Quality of Experience
SVR	Support Vector Regression
RF	Random Forest
ANN	Artificial Neural Network
SVM	Support Vector Machine
HOG	Histogram Oriented Gradients
GM	Gradient Magnitude
RO	Relative gradient Orientation
RM	Relative gradient Magnitude
SGD	Stochastic Gradient Descent

Contents

List of Tables

List of Figures

0.1 General Introduction

Vision is one of the most important ways for people to get information from the outside world. As humans see 3D scenes in nature, it has always been humans' goal to recreate accurate and natural 3D scenes on the screen.

In the beginning of the 10th century, several studies and remarkable work have been conducted by great physicists in the field of optics and light such al-Hassan ibn al-Haytham. The early scholar Ibn al-haytham had extensively affected the development of optics where he developed geometric optics theories and explained the visual perception mechanism. The scientist went deeper than anyone else in attempting to explain the fundamental physics of refraction and the structure of the human eye. He also introduced a camera pinhole box which further evolved into a 2D (Dimension) photographic camera in the first half of the 19th century. These enormous theories and researches are resulted in this series of books called *Book of Optics* [104]. After the evolution of photographic camera to make 2D analog images, the 3D media has started with David Brewster in 1844 where he created the stereoscope, the first portable 3D-viewing device. The pioneer in photography has also invented the binocular camera to be used in the stereoscope to take pictures. Years later, the stereoscopic camera became popular and changed the entertainment industry. The invention of semiconductor transistors in 1959 enabled the first appearance of a digital camera that was launched few years later.

Nowadays, with the rapid development of digital multimedia, 3D technology is taking the human's viewing experience to the next level. It gives the viewers with a more immersive and natural scene. This new wave of 3D media leads to an increase the expectation of the quality services especially in industry entertainment. According to the latest theatrical market statistics collected by the Motion Picture Association of America (MPAA), the number of worldwide 3D screens continued to grow in 2016 at a faster pace (17%) than in 2015 (15%) [86]. In the long run, the wide application of stereoscopic 3DTV broadcasting system is also expected. A pioneer for this was on 1 October 2010. The first High Definition (HD) stereoscopic 3D channel called SKY 3D, started broadcasting nationwide in United Kingdom at resolution of 1920 x 1080 pixels. The channel provided stereoscopic 3D contents including education, animation, sport, documentary and performances. However, the 3D content is not limited to the entertainment industry. 3D visualization concerns different applications, such as remote education [123], medical body exportation [122], robot navigation [8] and so forth. Therefore, it is reasonable to believe that the amount of 3D content will continue growing throughout the next few years.

0.1.1 Motivation

With the rapid development of 3D applications, 3D media on the Internet is increasing, and has been more and more widely used in peoples daily life and work. Therefore, perceptual quality assessment is critical factor in order to have a good viewing experience. In most of the 3D applications, the media quality could be affected by necessary treatments going from capture, compression, storage, transmission, to display. Since the 3D viewing provides immersive feeling, the contents quality needs to be assured in order to avoid any visual discomfort to the users. As a result, the requirement for precise and dependable objective image metrics has become even more important. However, compared to a 2D image, a stereoscopic 3D image involves depth information and consists of two views, which makes stereoscopic image quality assessment more challenging than 2D images.

0.1.2 Thesis aims and scope

The major objective of this thesis is to consider an industrial need for an objective stereoscopic image quality assessment, then, propose an accurate stereoscopic image quality methods to fulfil this need. It is important to account for human visual system (HVS) characteristics and properties (e.g., visual sensitivity, visual attention, and so forth). Therefore, in our proposed methods, we follow the idea of using human visual system modeling along with modern artificial intelligence techniques to judge the perceptual quality accurately. However, our research also investigates the quality assessment based on visual attention properties. Through this thesis, few questions are to be answered such as:

- What are the factors that affect the 3D viewing experience?
- How to objectively measure the perceptual quality of stereoscopic 3D images?
- What are the main aspects for designing stereoscopic image quality metric?
- Is the quality judgement in salient (human visual attention) regions?

The research topics covered several disciplines related to stereoscopic image quality assessment such as: Human Visual System (HVS), stereoscopic imaging, and machine learning algorithms. Overall, the thesis includes five chapters organized as follows:

Chapter One: introduces the concept of evaluating the quality of stereoscopic images. The chapter provides an overview of stereoscopic vision, the common types of artefacts, and general application of objective IQA metrics.

Chapter Two: introduces the most machine learning algorithms used for image quality assessment problem, and provides a state-of-the-art study of stereoscopic image quality metrics.

Chapter Three: introduces two proposed approaches for stereoscopic images, based on handcrafted quality features and HVS modeling. The first method is for measuring quality, while the second is for recognizing distortion types.

Chapter Four: In this chapter, we look at how machine learning techniques may be used to assess the quality of stereoscopic images and present two methods.

Chapter Five: This chapter investigates whether visual attention should be considered when designing an objective SIQA metric. To that purpose, we provide a novel metric that considers saliency information.

Chapter 1

Introduction to stereoscopic image quality evaluation

1.1 Introduction

Communication networks and digital technology advancement allows easy access through multiple devices (3D-TV, smartphones, tablets, etc.) for three-dimensional (3D) media contents with numerous applications such as education with virtual reality (VR), 3D navigation and 3D medical imaging analysis. This lead to rapid growth of 3D multimedia where stereoscopic imaging is usually used to display the 3D contents. However, the quality of these stereoscopic 3D contents can be affected at different necessary processing stages. The quality artefacts may occur during the dataflow of stereoscopic images as shown in Fig 1.1, from creation/capturing, compression/coding, transmission, decoding to display. Briefly addressed as follows:

- Capturing stage: Special care should be taken when placing cameras to avoid such blurry contents. The limitations of camera sensors can also cause noise to the captured pictures under particular circumstances. In stereoscopic imaging, disparities between left and right images provide depth information. Therefore, inadequate stereo camera configurations strongly influence the disparity and as a result, various artefacts may take place [53].

- Compression/Coding: Different 3D display systems rely on different 3D scene representation formats. Effective coding schemes of these data formats is key to the success of 3D displays and data storage. However, if the representation format is different from the one the scene was originally captured, converting between the formats is a source of quality degradations, where usually cause texture information loss [105]. In order to minimize transmission cost, redundant information is minimized by compression algorithms. Such algorithms are often improperly applied for stereoscopic contents, and binocular depth cues information may be lost in the process.

- Transmission stage: Packet losses of digital data can occur during the transfer through the network especially for wireless communications [49]. Resilience and error dissimulation algorithms attempt to minimize the packet loss, but if not designed for stereoscopic images, such algorithms might produce additional artefacts on their own.

- Decoding/Visual optimization: Some visual optimizations are required to cover the display limitations such screen resolution, aspect ratio, contrast range and so forth.

- Display: Several methods are established for 3D scene visualization, which offer distinct degree of scene approximation. Each family of 3D displays has its own characteristic quality degradations, in-which are often scene dependant [52].

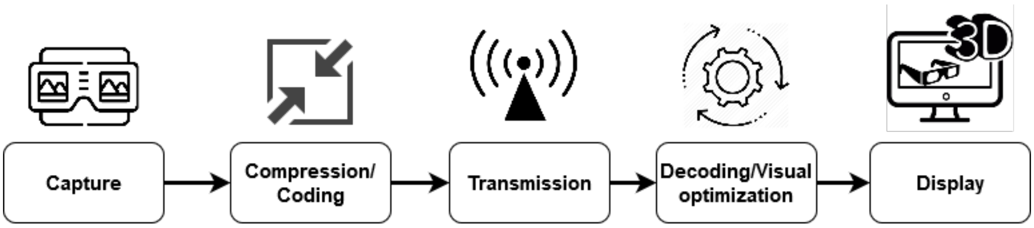


Figure 1.1: Illustration of different phases in stereoscopic image processing.

The aforementioned scenarios that decrease the stereoscopic image quality may cause visual symptoms to viewers such as headache, nausea, and visual discomfort/fatigue [63]. Thus, and in order to satisfy customers expectations for high quality imply determining the quality of stereoscopic contents reliably and effectively. An accurate Objective and subjective metrics are needed to evaluate the quality of stereoscopic contents. The subjective metrics are based on opinion scores given by human observers. While objective metrics are based on computer algorithms and numerical computations. However, in this thesis we strongly focus on objective methods.

The rest of the chapter discusses the basic mechanisms in relation to the Human Visual System (HVS) and the functional principles of simple stereoscopic systems, followed by a description of objective and subjective quality assessment for stereoscopic images.

1.2 Stereoscopic vision importance in imaging

Stereoscopic vision represents the ability of the visual brain to create a sense of three-dimensional structure and form from visual inputs. Therefore, using stereoscopic imaging technology can potentially improve many aspects of applications such as remote education [123], robot navigation [8] and so forth.

The stereoscopic imaging is also applied to medical body exportation [122]. For instance, researchers in [34] explored the benefits of using Stereoscopic Digital Mammography (DM) for breast cancer diagnoses. The authors have compared the standard 2D DM with the stereoscopic DM, where this latter showed higher diagnoses accuracy from doctors up to 90.9% versus 87.4% of 2D DM images.

Given the value of stereoscopic vision and its advantages on various applications, it is reasonable to assume that the amount of stereoscopic content will keep rising over the next few years.

1.3 Stereoscopic visual perception

HVS is still not fully understood. However, it is the one responsible for judging the perceived visual quality. This objective evaluation is reliable but not able to provide assessment for wide variety of stereoscopic content. This section introduces the basics of HVS and depth cues.

1.3.1 Human Visual System (HVS)

The HVS is indeed a combination of related subsystems that work together in a single process. Spatial, color and motion information is understood to be transmitted to the brain using mostly separate neural pathways. This visual pathway consists of a large number of nerve cells that transmit and receive electrical impulses. The visual system pathway begins from the eyes and proceeds to other parts of the brain that end up to the visual cortex cells. The wavelength of light is viewed and interpreted at various stages of the visual system as shown in Fig. 1.2. This interpretation starts from the retina of the eyes (Temporal/Nasal retina), going through the optic nerve, the optic chiasma to the Lateral Geniculate Nucleus (LGN), and finally the primary visual cortex. Almost all of the LGN connections go straight to the primary visual cortex. In this latter, elements are organized into series of rows and columns of neurons. The first neurons to obtain the signal are simple cortical cells. These cells detect lines at different angles, from horizontal to vertical, which occupy a large part of the retina [113].

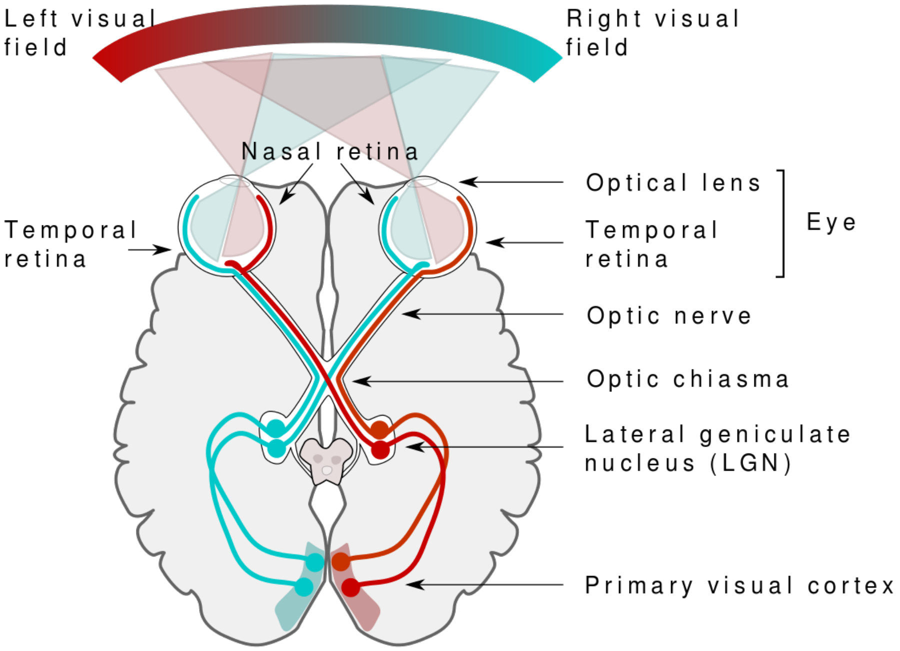


Figure 1.2: Human visual system pathway [88].

The HVS relies on various depth cues to reconstruct the 3D world from 2D images projected onto the retinal cells. These cues can be classified into three main groups [44]: 1) Oculomotor cues, based on the physical abilities of our eye muscles and lenses. 2) Monocular cues, where the HVS uses information from a single 2D view. 3) Binocular cues, that extract information from both eyes. This latter in particular has been simulated for the development of 3D imaging systems. The following section provides an overview of the binocular cues.

1.3.2 Monocular and Binocular depth cues

Monocular cues depend only on information from one eye to estimate depth. It extracts depth information from changes in retinal images over time (i.e. from movement). This type of depth perception is concluded from static 2D images and movement-based cues [44].

Binocular cues estimate depth information using the differences in the images received by the two eyes. Typically, the human eyes see the world from places that are around 6 cm apart. This shift in the point of view of the two eyes produces the cue of binocular disparity. Binocular vision derives to differences between left and right images projected to the left and right eyes. Thanks to this disparities, the HVS can form 3D scene and evaluate relative distances of objects. However, if the observer maintain his viewing position. The projections of an objects onto the left and right retina rely on the distance between the object and the viewer. Fig. 1.3 shows the observer fixates in a point F, which projects to the corresponding points F_L and F_R in the left and right retina, respectively. Meanwhile, another point P of an object, projects non corresponding points P_L and P_R onto the left and right retinas, respectively. Absolute disparity is defined as the difference in angular displacement between the projections of P and F. Noting respectively α and β as the angles between the projections of P and F onto the left and right retina. So the absolute disparity is given by the $\delta = \alpha - \beta$ [92]. However if we maintain the fixation F, the absolute disparity varies with the position and distance of point P from the observer, where the nearest object to the observer, the greater disparity δ and vice versa.

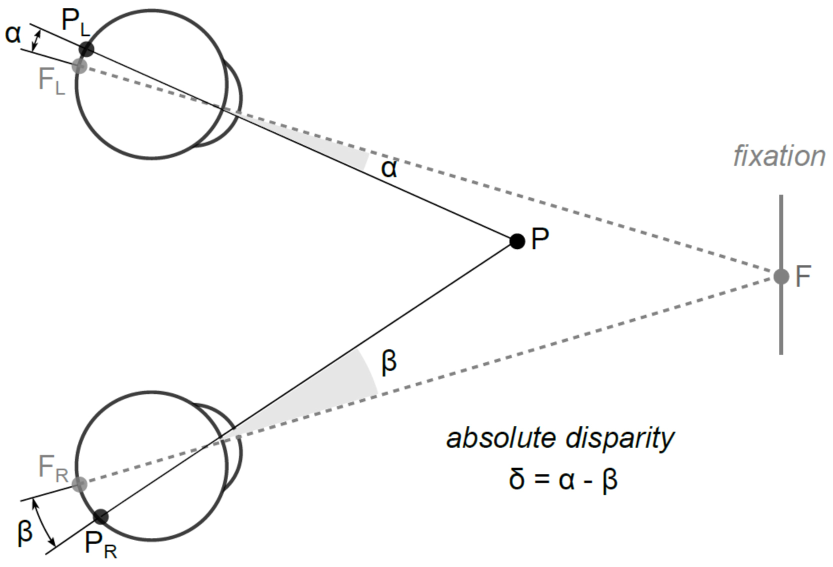


Figure 1.3: Basic geometry of human binocular vision [92].

1.3.3 Visual Discomfort

Visual discomfort occurs when eyes experience abnormal visual effects such as illusions of color, structure and motion. Given that the visual system has been adapted to the processing of natural images, the presence of illusions creates inadequate brain processing, which causes the viewer to experience pain, fatigue or exhaustion. However, in 3D visual experience, the distance between the eyes and the perceived object is called the focal length. Limited focal length would lead to visual pain, indicating symptoms such as eye strain, headache, fatigue, leading to unpleasant vision experience [64]. Therefore, in 3D stimuli such stereoscopic images, the amount of visual discomfort is primarily related to the depth information.

Although most studies on visual discomfort is mainly based on non-distorted images. Indeed, contents with degraded quality, provides nonuniform depth distribution that eventually may cause discomfort to the viewers. A good quality is then needed to prevent serious visual discomfort. In case of distorted 3D stereoscopic images, the reasonable assumption that can be made about the influence of visual quality has on visual discomfort. Is that people might be incapable of obtaining a stereopsis and estimating the depth, so the level of visual discomfort must be influenced. Besides this, studies in more details are required to investigate how strong is the relationship between visual discomfort and stereoscopic image quality.

1.4 Stereoscopic 3D imaging

The current 3D systems are based on the concepts of human depth of perception. According to binocular depth cues from binocular vision system, the HVS can interpret the world in 3D. This binocular depth perception is produced on the basis of slightly different locations of the two retinal images seen from the left and the right eyes. Also called binocular disparity which yields the perception of depth and inspires the development of 3D technology. In this section, we introduce the stereoscopic 3D vision for computers.

1.4.1 2D Images Acquisition System

Camera model called Pinhole in Fig. 1.4 is the simplest model considered to describe the formation of images/videos. In this model where C is the camera centre (pinhole) and f refers to the focal length. The images/videos are formed by projection on the image plane with center of P. When an image of a scene is captured by a camera, we lose depth information as objects and points in 3D space are mapped onto a 2D image plane as : $R^3, (x, y, z) \longrightarrow R^2, (x, y)$.

This 2D Image/video Acquisition is dependent on sensors inside of camera to form pictures. These sensors are usually placed behind the center of the camera C with the same distance f. They convert the light rays to electrical charges and typically presenting them with RGB (Red, Green, Blue) color model. Commonly, there are two different technologies to represent each physical point M from the scene to pixel/image element: CCD (charge coupled device) and CMOS (complementary metal oxide semiconductor). However, each type has unique strengths and weaknesses giving advantages in different applications [26]. A point in the 3D world $\mathbf{M} = [X, Y, Z]^T$ is then mapped to $\mathbf{m} = [x, y]^T$ on the image plane according to the following relationship:

$$x = f\frac{X}{Z} \qquad y = f\frac{Y}{Z} \tag{1.1}$$

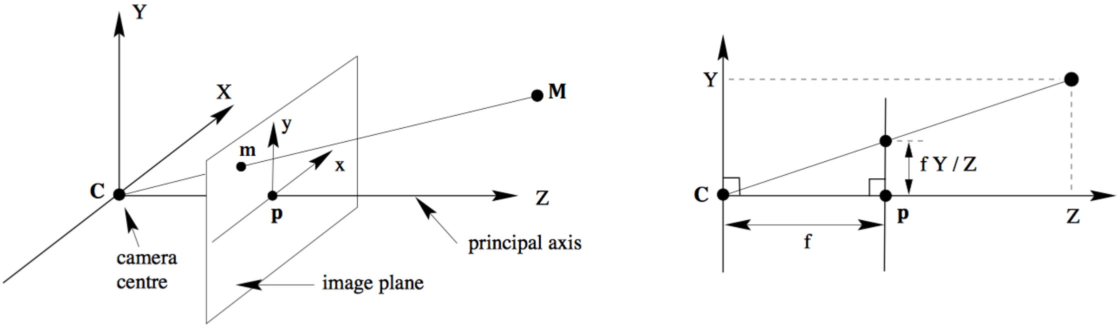


Figure 1.4: Basic geometry of Pinhole Camera model.

1.4.2 3D Stereoscopic Images Acquisition System

The process of capturing stereoscopic images is an effort to mimic what we see through our two eyes. The basic concept for imitating the HVS is therefore to replace the left and right eyes with two horizontally separated cameras. While in display, the concept uses a screen that projects the left and right views to the respective eyes. Then the brain fuses these images, resulting in a deep perception. Currently, the most widely used stereoscopic camera system is seen in Fig. 1.5.

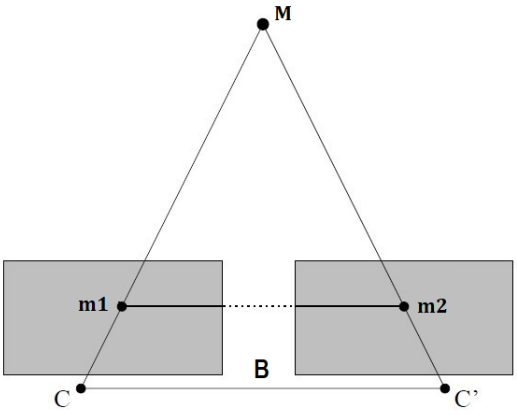


Figure 1.5: Pinhole model of stereoscopic camera systems.

Where C and C' are the camera centers, B refers to the baseline (distance between the camera centers), and M is a 3D point projected to the left and right 2D plans giving $m1$, and $m2$ respectively. It simulates the human binocular disparity. However, a calibration techniques are necessary to align pixel information between the cameras and extract the lost depth information in the projection process. The calibration problem is often considered solved, but recent research still focuses on the subject because it directly effect the quality of 3D reconstruction from the stereoscopic images [43, 90]. It consists in finding the internal and external geometry of the acquisition system such the focal length, the optical center of the camera, the dimensions of the pixel, the angle of obliquity of the pixel, and so forth.

1.4.3 Stereoscopic 3D displays

Stereoscopic 3D display is needed to visualize 3D images/videos. The technology behind 3D displays has strengths and limitations in the production of high quality 3D content. The different display types also influence the viewer's quality of experience. As listed in table 1.1, there main categories are denoted to distinguish the 3D displays: (1) direct view stereoscopic displays, which require eyewear and classified based on the multiplexing method; (2) auto-stereoscopic direct-view displays; (3) binocular head-mounted displays, which the stereoscopic projections are integrated into the eyewear device itself and thus do not require glasses. [114].

Table 1.1: Stereoscopic 3D displays classification.

Categories	Stereoscopic direct-view (require glasses)	Auto-stereoscopic direct-view (no eyewear)	Head-mounted and interactive (wearable)
Technology types	• Color multiplexed • Polarization multiplexed • Time multiplexed (shutter glasses)	• Two-view • Multi-view • Head tracked • Light field	• Optical head-mounted projection (e.g virtual reality applications)

- Stereoscopic direct-view visualisations require the observer to wear glasses to direct the left and right images to the relevant eye.

- Auto-stereoscopic displays do not require any glasses to present two-view images, but send them directly to the corresponding eyes using aligned optical elements on the surface of the displays [30]. This type of displays simplifies the viewer's 3D experience and can display multiple views, making 3D entertainment more applicable. This displaying approach projects each view from a specific viewing angle along the horizontal direction and provide a comfortable viewing zone for each stereoscopic image.

- Head-mounted monitors are binocular systems where it usually consists of two separate mini displays with connected relay optics. Since the gadget is carried on the head, the user is not bound to a fixed viewing location and can perceive complete immersion from the viewing scene [94].

As listed in Table 1.1, there are three types of 3D eyeglasses that correspond to the three ways stereo frames are separated for 3D effects: anaglyph, polarized and Time multiplexed. Samples are showed in Figure 1.6.

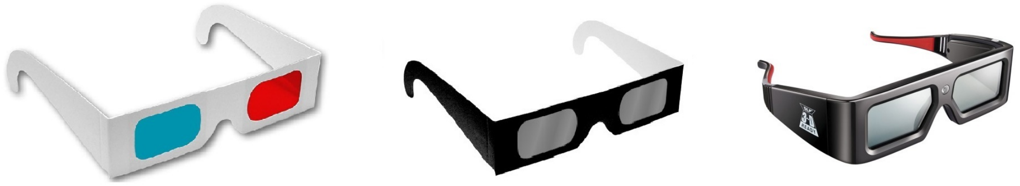


Figure 1.6: Pictures of the three different types of 3D glasses. Left: Color multiplexed (anaglyph); Middle: Polarization multiplexed; Right: Time multiplexed (shutter glasses)

- Anaglyph: Based on color filters; Red/cyan, red/blue and red/green glasses are available in paper and plastic frames. The glasses would allow one color to flow through one eye while blocking the other. This ensured that both of our eyes viewed the two distinct images that our brain recognized as 3D.

- Polarized: The most popular among the other two. It is based on the concept of linear polarization. One lens carries a vertical linear polarizer and the other would have a horizontal linear polarizer. This ensured that both the eyes had a different image for the brain and that the original colour of the image is maintained unlike Anaglyph glasses.

- Shutter Glasses: Modern glasses that works by only presenting the image intended for the left eye while blocking the right eye's view, then inverting this process for the other eye, and repeating this in time of milliseconds. Unlike polarized glasses, where the horizontal spatial resolution is normally reduced, the active shutter system can keep full resolution for both the left and right pictures.

In addition to other quality degradations inherent in the acquisition of stereoscopic images. The technologies listed above are not exempt from the flaws that cause visualization artefacts. However, when two mechanical projectors are used to present an image to each eye, misalignment problems are common issues and potentially occur while displaying. Another common artefact inherent to stereoscopic visualization is caused by poor image luminance and contrast due to light losses in filter-based systems and glass systems. However, the study in this thesis will not focus on stereoscopic 3D displays and their artefacts on the perceived stereoscopic contents.

1.4.4 Stereoscopic disparity and depth map

The depth map denotes the distance between the objects of the scene and the viewer's point of view. Disparity map refers to an image containing the distance between two respective pixels in the left and right views of the stereo pair. However, a depth map can be estimated using a 2D image, while a disparity map can only be obtained using a stereopair image. Note that the disparity value can be translated to a depth value based on a particular formula and vice versa. Nowadays, disparity/depth maps are important in many applications, such as augmented reality, 3D reconstruction and navigation.

The disparity information is proven to be a strong effective factor for stereoscopic images and videos quality, where researchers first expanded 2D Image Quality Assessment (IQA) metrics to Stereoscopic Image Quality Assessment metrics (SIQA) by adding the analysis about the depth information (Benoit et al [12]). Since most of the artefacts directly impact the disparity/depth information. Currently, this latter has become necessary for assessing the quality of the stereoscopic content.

In a stereoscopic images, the depth of an object in the 3D space is related to the difference of its appearance in the left and right view. This disparity is presented in the stereoscopic image by shifted pixels horizontally or vertically between the left $L(x, y)$ and right $R(x', y')$ pixels that correspond to the object. As demonstrated in Fig. 1.7, let $L(x, y)$, the position of left pixel and its corresponding pixel on the right view be $R(x, y')$. The pixel that reflects the same object is shifted horizontally on the right image by N number of pixels, where $N = y - y'$. So for this horizontal disparity, we denote the disparity map as $Di(x, y)$, where values of this matrix represent the number of pixels referring to the distance between the left pixels and its corresponding ones on the right. The map can then be computed as follows:

$$Di(x, y) = y - y' \tag{1.2}$$

Measurement or assessment of depth and disparity is essentially the same concept. We denote a matrix $De(x, y)$ that refers to the number representing how far (depth) the object is from the camera. The depth map is defined as follows:

$$De(x, y) = \frac{f \times B}{Di(x, y)} \tag{1.3}$$

Where f is the focal length and B is the baseline (distance between the two cameras, see Fig. 1.5). However, the estimation of depth/disparity maps has always been a challenging task. The diversity and complexity of objects in the scenes makes the estimation difficult to obtain accurate map as the ground-truth data. Most of disparity estimation schemes follow the three steps to locate each pixel on the left view $L(x, y)$, its corresponding pixel on the right view $R(x', y')$ is required to be:

- (1) On the same row: $x' = x$;
- (2) To the left of $(x, y) : y' > y$;
- (3) Most similar to the pixel $L(x, y)$ among all candidates found after the previous two steps. In order to determine the most relevant pixel, the most widely used approaches are based mainly on a block matching strategy.

For instance, a method that compares the sum of absolute value (SAD) of the neighbors of the pixel [60]. Another method called a semi-global block matching suggested in [51], where the authors improved the accuracy similar using more neighboring blocks. More recently in [21], the authors employed structural similarity (SSIM) [120] for this purpose and reported its superiority than SAD. Figure 1.8 shows the left (a) and right (b) images, the depth map (ground-truth) (c) and disparity map (d). The disparity map is estimated using the scheme from [21]. Overall, the disparity map is well estimated with some flaws at some regions.

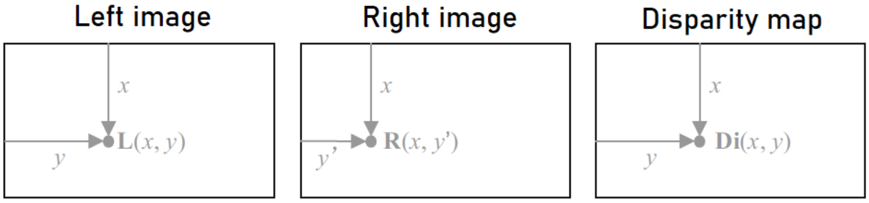
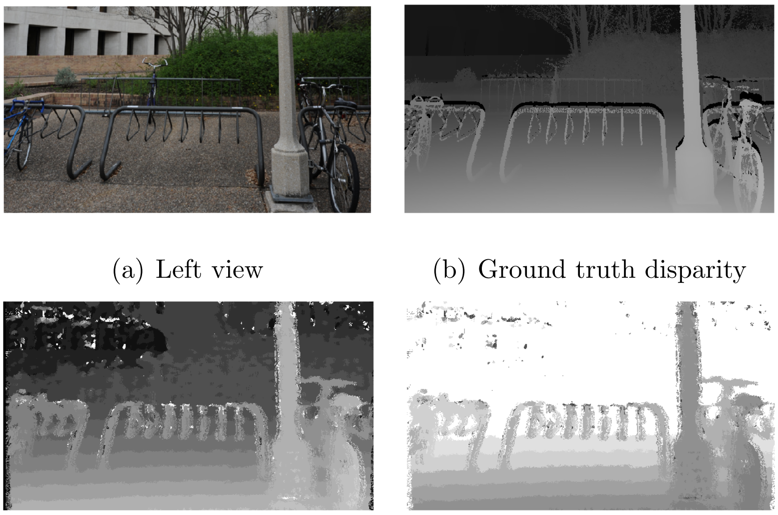


Figure 1.7: Left and right view pixels and its correspond disparity.



(a) Left view (b) Ground truth disparity

(c) Estimated disparity (d) Estimated depth

Figure 1.8: An example of disparity and depth map estimation from stereoscopic image [82].

1.4.5 3D reconstruction

Due to the numerous applications, the 3D scene reconstruction, which aims to represent a scene in three dimensions, is receiving a lot of interest recently. While the researchers focus is on the acquisition of very high quality three-dimensional media.

Essentially, there are two approaches for 3D data acquisition. On the one hand, the active methods, they acquire the depth of a scene from a controlled light source such as laser beams. On the other hand, the passive methods which are based on computer vision algorithms. Where the 3D acquisition is based on a set of images of the scene. The second method is mostly used for its simplicity and cost-free. Some approaches use only one image such as the shape-from-shading [33]; others, such as stereophotometry, exploit several images taken from the same angle and different illuminations [32]. For that, stereoscopic imaging are commonly used for the reconstruction of a 3D scene. As a result, it is reasonable to validate the quality of the stereoscopic media before the 3D reconstruction.

However, there is another method that lies between the active and passive methods, the structured-light 3D scanner that uses both of structured (active) light and images of the scene: light patterns are projected on the scene at the moment of image acquisition creating an additional texture on the surface. Despite of the high performance of this system, it still requires the purchase of expensive equipment.

1.5 Common quality degradations

As discussed earlier in section 1.1, the quality of stereoscopic image general can be affected by many factors due to the necessary treatments (acquisition, processing, compression, transmission, etc). Quality estimation is therefore required and can be a key factor in the design and optimization of stereoscopic image content delivery systems. The first step towards objective quality estimation metric is to identify the artefacts which could arise when dealing with stereoscopic content. In this section we denote the most common degradation types of stereoscopic images which include: Blur, Blocking, noise, resizing, and contrast.

1.5.1 Blur

Blur's distortion affects the edges of an objects in the images. This distortion makes the objects unrecognizable and difficult to perceive. Blurring is defined by the loss of the high frequency information present in the images [5]. This phenomenon smooths the image signal and with higher blurring the smoother signal is, which causes the reduction of signal edge points as described the example in Fig. 1.9. However, blur distortions can be present in different forms denoted as follows:

- Defocus: Also known as out-of-focus and occurs during capture. To make sure that an object is sharply mapped on the sensor, the focus of the camera must lie within the depth of field. The size of this area depends on several parameters such focal length, lens aperture, and the object distance. Therefore, objects outside this area are defocused and appear blurred. The inappropriate focus on each object leads to the blurring of the entire scene. Making attention to the shooting conditions and changing the camera parameters correctly is the best way to prevent this kind of distortion.

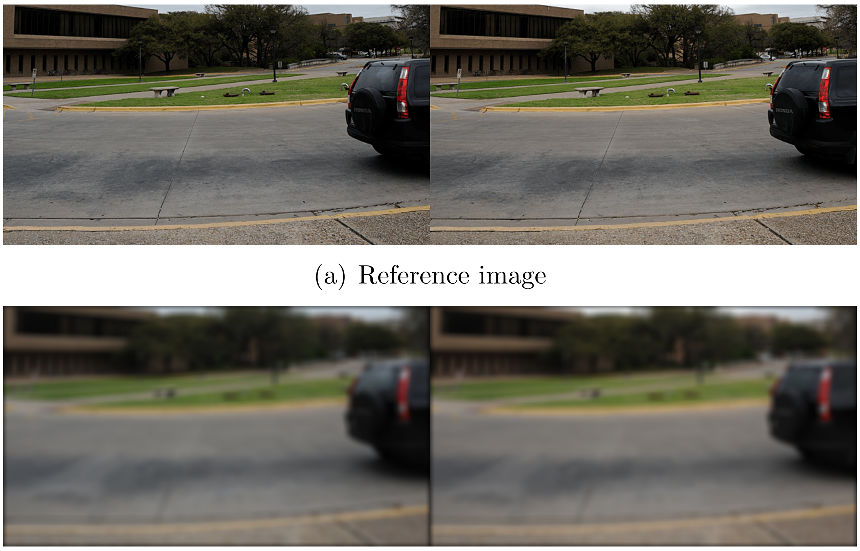
- Motion blur: Refers to the blur caused by the rapid movement of objects photographed during recording. Motion blur is formed either by the movement of the object when the capture device is stationary, or by the scene to be filmed if the camera follows the moving object. This type degrades specific directions in the frequency domain.

- Blur due to Compression: The two well used compression algorithms called Joint Photographic Experts Group (JPEG, and JPEG2000) [42] are an important source of blur. In general, low pass filters are applied to the image, they operate at high frequencies. This induces a loss of details and sharpness.

- Processing blur: During the processing phase, filters can be applied to the image, and thus may develop quality degradation. Filtering is often responsible for blurry distortions.

- Transmission blurring: The transmission of images in the channels often produces loss of information, these losses result in blurred regions. The manner usually occur in wireless communication Rayleigh channels.

Due to the blur distortion, the depth information may not be derived correctly from the stereoscopic image. There are several approaches available to restore the focus of



(a) Reference image

(b) Distorted image

Figure 1.9: Blur distortion effect on stereoscopic image [82]. (a) Reference image without distortion, (b) image with Blur distortion.

a blurred picture via inverse filtering, but most of them cannot be used practically in real-time applications [16].

1.5.2 Ringing effect

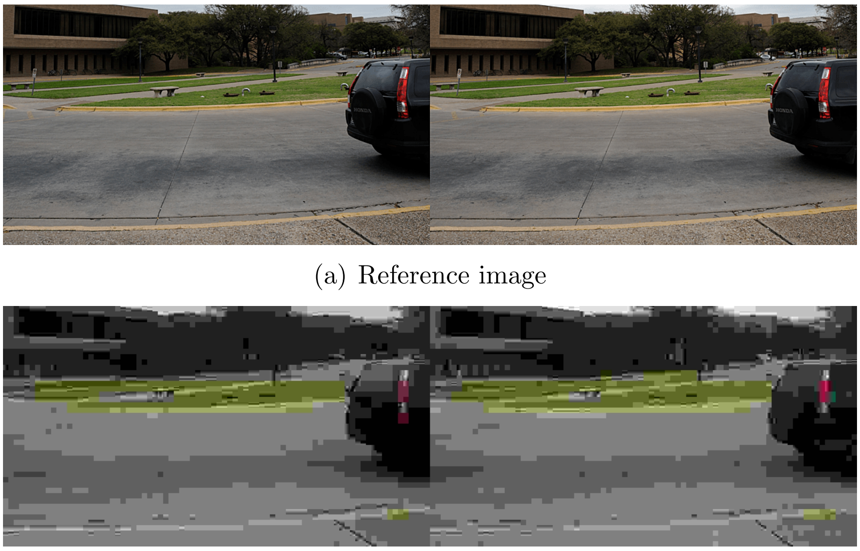
Ringing effect can be introduced to oversharpened images, images transmitted over analog channel, or after image processing algorithms such compression. In compression algorithms, this degradation is generally due to the step of quantization or decimating the high frequency coefficients [37]. It manifests in the form of oscillations on high contrast regions and is often defined as noise around these regions.

This phenomenon appeared as rippling artifact near sharp edges of the image. Indeed this artefact is annoying to the observers especially for stereoscopic stimulus.

1.5.3 Blocking effect

Block-artifacts are a result of block transform coding. Where this transform is a common process in JPEG/JPEG2000 compression. In 2D/3D stereoscopic images this distortion is expressed at the boundaries between blocks and appears as vertical and horizontal contours [98]. However, any lossy block-based coding scheme introduces visible artifacts in pixel blocks and at block boundaries. These boundaries can transform block boundaries, prediction block boundaries, or both. The transform (for example the discrete cosine transform) is applied to a block of pixels, and to achieve lossy compression, the transform coefficients of each block are quantized. The lower the bit rate, the more coarsely the coefficients are represented and the more coefficients are quantized to zero. Since this quantization process is applied individually in each block, neighboring blocks quantize coefficients differently, which causes discontinuities at the block boundaries.

In images that have more low-frequency than high-frequency content, the low-frequency content remains after quantization, which results in blurry, low-resolution blocks. The blocking effects are visible in the stereoscopic image shown in Fig.1.10. The stereo image presents an imprecise contour which causes degradation of visual aesthetic quality. This type of degradation usually occurs on blurred images, whose edges of objects are more diffuse.



(a) Reference image

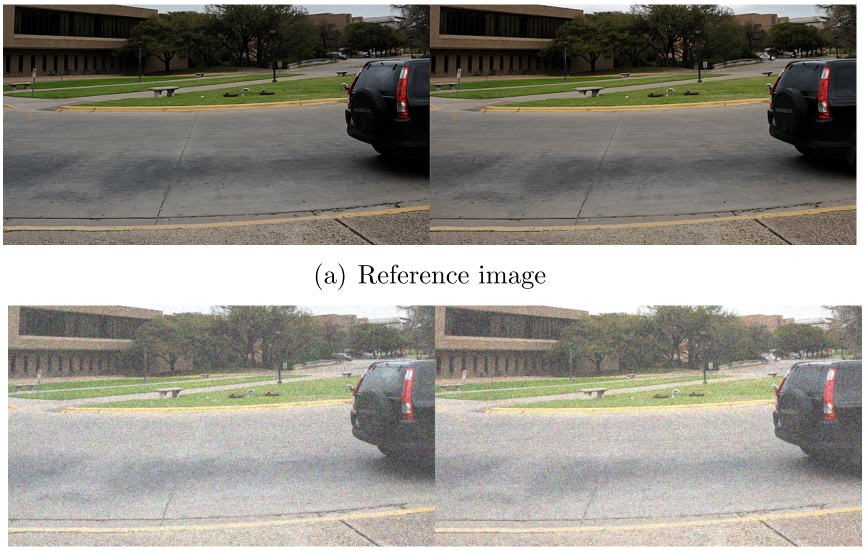
(b) Distorted image

Figure 1.10: Blocking effect from JPEG compression on stereoscopic image [82]. (a) Reference image without distortion, (b) image with JPEG distortion.

1.5.4 Noise

Noise is a common degradation of images. Any parasitic information added to the picture is described as noise [38]. Mostly characterized by the presence of visible grains which

causes visual discomfort and annoys the viewers. Noise has different forms based on the origins of this artefact such as temporal and spatial noise, these latter could be presented to the stereoscopic images. Noise has relation with electronic components. The CCD camera sensors introduce noise at some circumstances such as high temperature.
However, to mimic the effects of this artefact, an Additive white Gaussian noise (AWGN) is a basic noise model denoted by specific characteristics. Stereoscopic contents are highly exposed to noise, Fig. 1.11 shows stereoscopic image with AWGN. Additive because it is added to any noise that might be intrinsic to the information system. White refers to the idea that it has uniform power across the frequency band for the information system. It is an analogy to the color white which has uniform emissions at all frequencies in the visible spectrum. Gaussian because it has a normal distribution in the time domain with an average time domain value of zero.



(a) Reference image

(b) Reference image

Figure 1.11: Stereoscopic image with white Gaussian noise distortion [82]. (a) Reference image without distortion, (b) image with white Gaussian noise distortion.

1.5.5 Upscale/Downscale

Display monitors are available in different range of screen sizes, requiring up-scaling or down-scaling of the content to match the screen resolution. Interpolation and decimation algorithms have been designed to change the spatial resolution. However, a loss of information and spatial details is unavoidable in the case of decimation, and a trade-off is always expected. Due to imperfect re-sampling algorithms, this issue causes blurring artefact and reduce spatial information captured of the scene.

1.5.6 Contrast

Contrast is the variation in luminance or color that makes an object in the scene distinguishable. It is determined by the imbalance in color and brightness of the object and other objects within the same field of view. However, the human visual system is more sensitive to contrast than to absolute luminance. Regardless of the changes in illumination over the day or from place to another, we can perceive the world similarly.

Contrast is a determining factor in the perception of visual quality [132]. A bad contrast is often occurred during the acquisition phase. Hardware limitations acquisition conditions and lighting conditions are the main causes of loss of contrast and visibility of scene details. There are other sources of contrast distortion, among these sources include the enhancement process that creates loss or over-contrast [16].

1.6 Distortions impact on the disparity/depth map

The distortions addressed earlier influence the 2D images and stereoscopic images quality differently because the degradation will have direct impact on stereoscopic image disparity that carries depth information. Since that the human brain uses disparity information to see the world in 3D. The disparity information is proven to be a strong effective factor for stereo images/videos quality judgment [18]. For instance, If depth or disparity is not properly available, the viewer perceives wrong distances between objects in the scene. Figure 1.12 illustrates how three types of distortion impact the estimated disparity map using the same stereo matching algorithm. Taking color range from dark to white, the closer object to the camera the lighter color is. As can be seen, each distortion impact differently disparity map. Blur distortions tend to cause disparity losses in particular for far objects, while JPEG/JPEG200 compression artifacts cause arbitrary loss related to blocking effect.

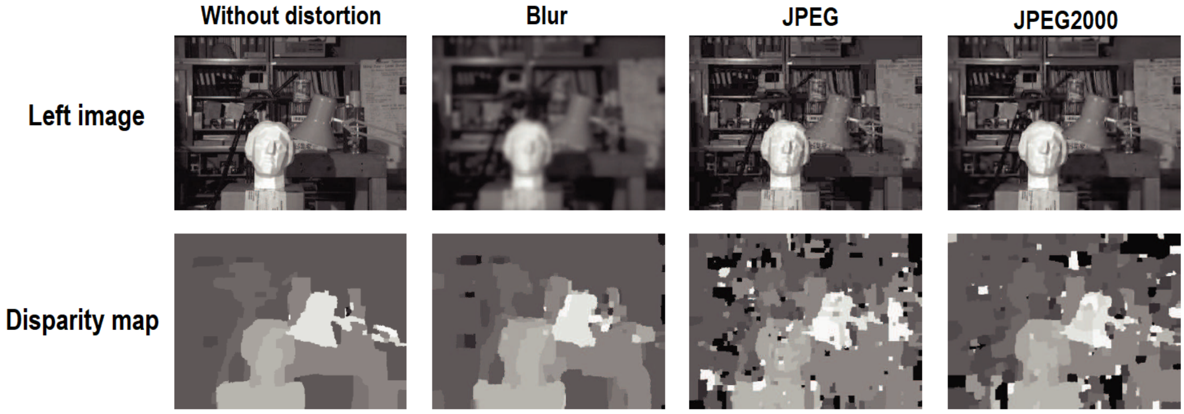


Figure 1.12: Disparity maps estimation from stereoscopic images under different distortions [18].

1.7 Binocular rivalry

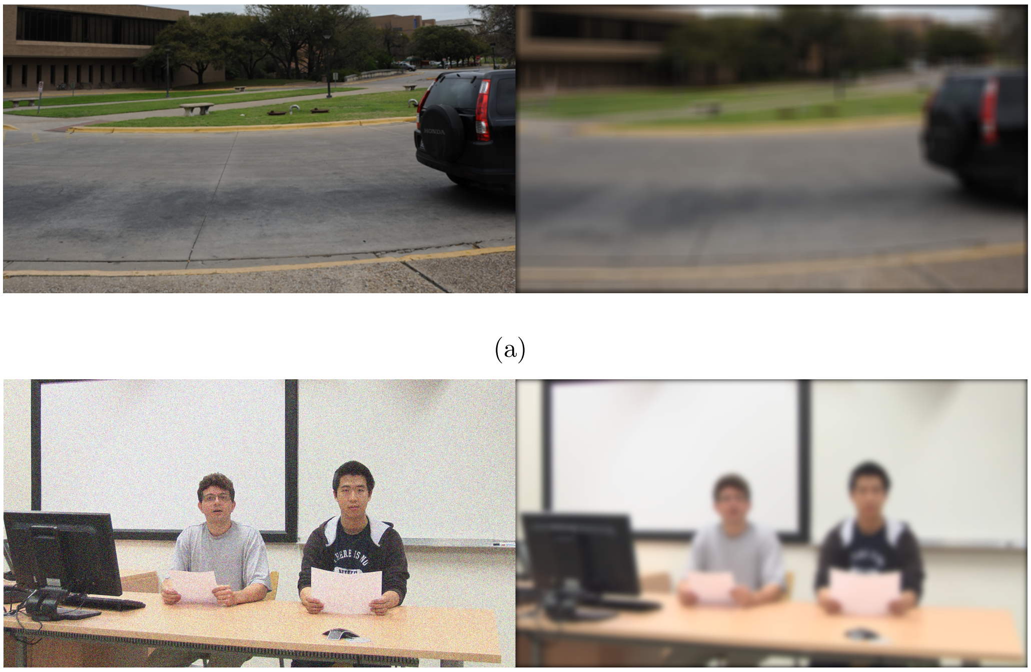
While the left and right images are consistent, they are fused in the visual system to a single percept of the scene, known as binocular fusion. Binocular rivalry is a phenomenon of visual perception in which perception alternates between different images presented to each eye. This could be very annoying to observers and usually causes fatigue and headaches. However, an asymmetric distortion (in section 1.7.1) in stereo contents is one of the cases that causes this phenomenon.

To address the question of where in the brain rivalry occurs, Blake *et al* [14] have studied Neural responses in the LGN and the visual cortex (as discussed earlier in section 1.3.1). Where they conducted several electrophysiological experiments, in-which binocular rivalry stimulus has been used. The authors concluded that in species with well-developed binocular vision such as Humans and monkeys, the retinal terminals from each eye project to different layers in the LGN, so that they remain segregated. Each layer receives excitatory input from one eye and contains a detailed retinotopic map of the contralateral visual field. The maps are in perfect register and receive feedback from primary visual cortex, which can detect mismatches in visual attributes such as orientation, spatial frequency or direction.

Binocular fusion mechanism and binocular rivalry provide a potential theory to develop 3D quality prediction models. Although there is a rich literature on binocular fusion and rivalry in neural vision science, simulating and applying the concept to stereo IQA remains an active research area.

1.7.1 Asymmetric distortion problem

The stereoscopic contents are possible to be distorted in asymmetric way, it is particular case of binocular rivalry where left and right views have distinct deformation. This makes objective Stereoscopic IQA problem more challenging than 2D IQA. However, there are three possibilities for asymmetric distortion case: 1) One view is distorted and the other is not. 2) The two views are affected with different type of distortion. 3) Both views are distorted by the same type of deformation, but with various degrees. Figure 1.13 illustrates examples of the first and second scenarios.



(a)

(b)

Figure 1.13: Asymmetric distorted stereoscopic images: (a) left-view is original and the right-view is blurred [21]. (b) Left-view has white noise, right-view is blurred.

1.8 Subjective quality assessment

Subjective quality assessment is relied on human ratings, where observers give their opinion on perceived contents. There are various subjective quality evaluation methodologies, and they are reliable to measure the quality of experience of any multimedia service for a user. But this type of assessment are typically time consuming and require a large number of users to produce accurate results. In subjective assessments, a group of human observers is asked to evaluate the quality of stimuli that are presented according to a specific procedure. The composition of the group can vary from one application area to the other. It is mostly desirable for the panel to cover as wide range with respect to age, gender, and cultural background as possible. In order to obtain accurate and reproducible subjective results, it is important to identify and explain each of the following elements:

- Laboratory equipment, including details of monitors and their arrangement, screen and viewing distance, illumination and characteristics of the room.
- Data set, including the original and after processing contents and their distribution across different sessions.
- Test methodology, including the rating target (quality, comparison, or impairment), the scale (categorical or continuous) and the type of stimuli.
- Score processing, including score normalization, outlier detection, mean score and confidence intervals computation, and significance tests.

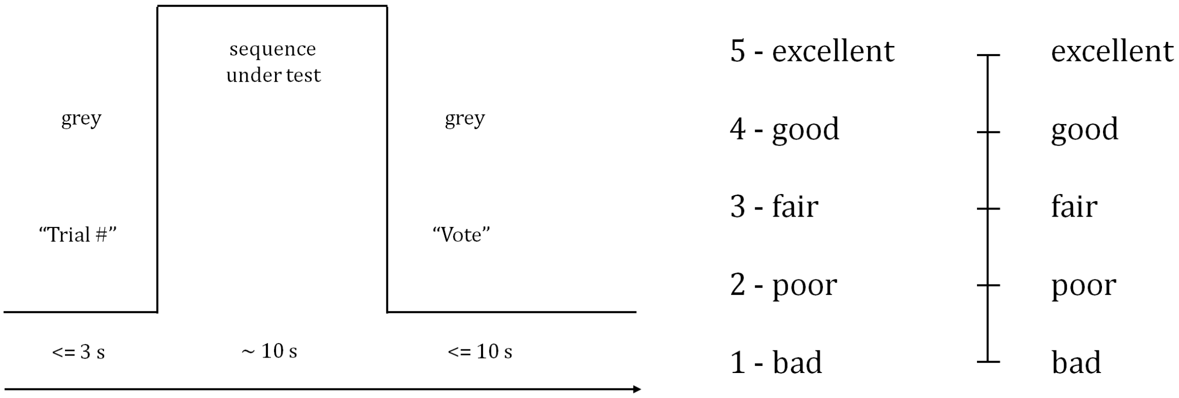
However, a subjective online QoE evaluation framework may be deployed. Recently after the COVID-19 pandemic, engineers and dedicated laboratories prefer to deploy more of this system, where it can be difficult for 3D stimuli due to the limitations of test equipment from the online observers.

1.8.1 Subjective assessment protocols

Subjective methods are much dependent on the nature of the test and can be substantially biased when not carefully planned, performed and interpreted. In order to maximize the reliability and reproducibility of the experiments, number of standards and recommendations has been issued by the International Communication Union (ITU). A group of experts described the conditions, procedures, processing of results [1, 3, 2].

The procedures for subjective tests can be divided into four scaling methods, we discuss them as follows:

- Absolute category rating (ACR): Also known as single stimulus method, this rating method is a category judgment where the test sequences are presented one at a time and are rated independently on a category scale. The method specifies that after each presentation the subjects are asked to evaluate the quality of the sequence shown. The time pattern for the stimulus presentation can be illustrated by Fig. 1.14. In the case of using a constant voting time, then the voting time should be less than or equal to 10 seconds. The presentation time may be reduced or increased according to the content of the test material. After the stimuli is prepared, during the presentation the observer has to judge it by selecting a discrete rating or giving continues rating as shown in sub Fig. 1.14 (b) and (c).



(a) The structure of a trial assessment. (b) Discrete scale. (c) Continuous scale

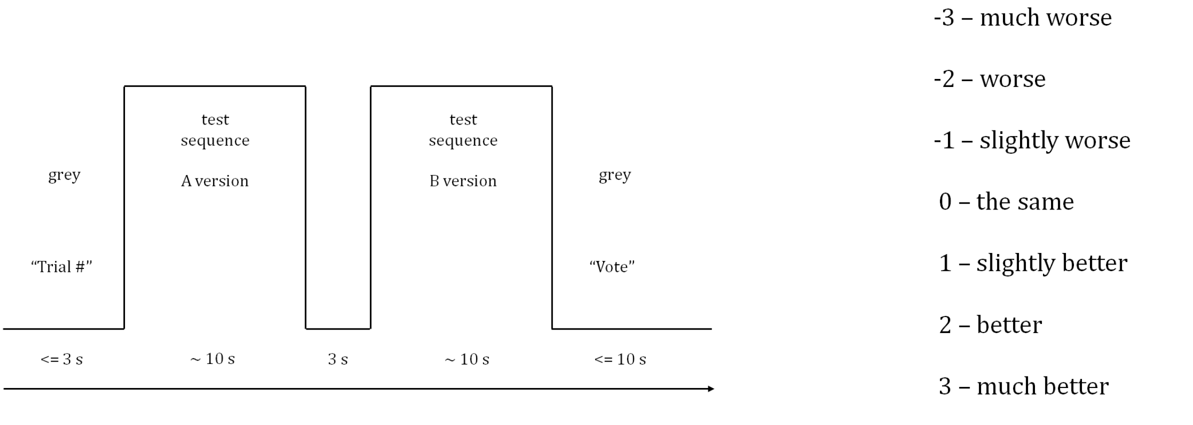
Figure 1.14: (a) The structure of a trial assessment proposed for the evaluation of stereoscopic images. (b) and (c) The labeled discrete and continues five-rating ITU scale for the subjective assessment of stereoscopic image quality [1].

- Absolute category rating with hidden reference (ACR-HR): This rating a judgment with hidden reference where the test sequences are presented one at a time and are rated independently on a category scale. During the data analysis, a differential quality score (DMOS) will be computed between each test sequence and its corresponding (hidden) reference. This procedure is known as "hidden reference". The method specifies that, after each presentation, the subjects are asked to evaluate the quality of the sequence shown. However, if a constant voting time is used, then the voting time should be less than or equal to 10 seconds. As in the ACR rating, the five-level scale for rating overall quality is used.

- Degradation category rating (DCR): Also called the double stimulus impairment scale method. The degradation category rating implies that the test sequences are presented in pairs: the first stimulus presented in each pair is always the source reference, while the second stimulus is the same source presented through one of the systems under test. As in the previous rating categories, the voting time should be less than or equal to 10 seconds if a constant voting time is used.

- Pair comparison method (PC): This protocol, also referred to as the paired comparison (PC) method, consists of a series of trial assessments during which the participants needs to compare two images displayed simultaneously, preceded and followed by mid-gray displays, exactly as in Fig. 1.15. The two stimuli can be also presented sequentially with 3 seconds mid-gray display between them. The number of trial assessments needed in one experiment is the one that covers all the combinations of any two such stimuli. Since the judgments in this protocol are in terms of preference, they can be expressed either using a binary scale, or by giving a graded preference on a scale as shown in sub Fig. 1.15 (b).

At the end of the experiment, the individual opinion scores can be concentrated into mean opinion scores (MOS). However, if reference images are included in the test sessions, the difference opinion scores between the scores of the distorted images and the scores of their corresponding references can be calculated, then the difference mean opinion scores (DMOS) obtained [2].



(a) (b)

Figure 1.15: (a) The structure of a trial assessment with a sequential presentation of the two stimuli for the SC method [1]. (b) The labeled ITU graded scale for the subjective assessment of (stereoscopic) image quality with the SC method [1].

1.8.2 Stereoscopic image quality databases

Subjective assessments are an important tool for building IQA databases, which include images with various forms of distortions and subjective opinions for all images, whether in form of MOS or DMOS. The past twenty years, the quality management community has known several publicly accessible stereoscopic 3D-IQA databases. In the following, we cite six most popular stereoscopic IQA databases:

- **IRCCyN/IVC 3D [12]** : The IVC 3D Image Quality Database has been established in 2008. It is the first public-domain database on stereoscopic image quality. Test conditions include JPEG and JPEG2000 compression as well as Blur. This dataset contains 96 stereoscopic images and their associated subjective scores. The resolution of these images is 512 × 512 pixels. 6 different stereoscopic images are used in this database which is composed of 6 reference images and 16 distorted versions of each source generated from 3 different distortion types (JPEG, JP2K, Blur) symmetrically to the stereopair images.

- **LIVE 3D phase I [82]** : The phase I consists of 365 distorted stereo images with a resolution of 640 × 360 pixels. There are eighty stereo images for each JPEG, JPEG2000 (JP2K), White Noise (WN), and Fast Fading (FF). The remaining 45 stereo images represent Blur distortion. All distortions are symmetric in nature. The subjective evaluation scores are given in the term of DMOS within the range of [-10,70].

- **LIVE 3D phase II [21]** : Phase II consists of 360 distorted stereoscopic images. This database includes asymmetric and symmetric distorted stereopairs over five types of distortion as the phase I. Specifically, 120 stereopairs are symmetrically distorted and the rest 240 stereopairs are asymmetrically distorted. The two phases constitute the largest and most comprehensive stereoscopic image quality database currently available. The three publicly available datasets have been used to test the performance of the proposed model on several different types of distortion. Subjective evaluation scores are given within the range of [20,80] in the DMOS term.

- **Waterloo IVC 3D Phase 1 [117]**: It has 330 full HD (1920 x 1080 pixels) distorted stereo images derived from six pristine stereo images collected from the Middlebury Stereo 2005 data sets. Three forms of distortion are present in this database: additive white Gaussian noise, Gaussian blur, and JPEG compression. These distortions are performed symmetrically on 180 stereoscopic images and asymmetrically on the rest 150 stereopairs. Subjective evaluation scores are given in term of MOS and distributed in the interval of [10,100].

- **Waterloo IVC 3D Phase 2 [116]**: It contains 460 full HD stereo images created from 10 pristine stereo image pairs. The stereo images carry the same distortion types as Phase 1, and both of them include symmetric and asymmetric distortions. In this database, 210 stereoscopic images are distorted symmetrically, and the rest 250 stereoscopic images are distorted asymmetrically. Subjective assessment scores are in term of MOS and the range is the same of Waterloo-P1 ([10,100]).

- **MCL-3D [106]**: The called MCL-3D database has 693 stereoscopic image pairs, where 1/3 of them are of resolution 1024x728 and 2/3 are of resolution 1920x1080. Gaussian blur, additive white noise, down-sampling blur, JPEG and JPEG-2000

(JP2K) compression and transmission error are the distortion forms added to either the texture image or the depth image before stereoscopic image rendering. The pairwise comparison was adopted in the subjective test and the Mean Opinion Score (MOS) was computed accordingly.

It is worth noting that the asymmetric degradations in the Waterloo phase 1 and 2 databases are different from those in the LIVE phase II database. This latter uses only one type of distortion to perform the asymmetry, while the two Waterloo databases consider the possibility of multiple types of degradation in which the left and the right images are affected by different distortions. All six of the above databases are publicly accessible. The creation of an IQA database is expensive and time-consuming, thanks to the researchers who provided these data sets. However, there are also non publicly available SIQA databases such as: NBU 3D I [95], MICT 3D [126], and SVBL 3D [118]. Table 5.2 summaries the discussed SIQA databases.

Table 1.2: Summary of stereoscopic IQA databases. Sym and Asym denote separately the symmetric and asymmetric distortion. R.S refers to reference scenes, while P.A refers to Publicly Available database.

Database	R.S	Resolution	P.A	Sym./Asym.	Depth map	Distortions
IVC 3D [12]	90	512 x 512	YES	YES/NO	NO	JP2K, JPEG, Blur, down/up scaling
3D LIVE P-I [82]	20	360 x 640	YES	YES/NO	YES	JP2K, JPEG, WN, Blur, FF
3D LIVE P-II [21]	8	360 x 640	YES	YES/YES	YES	JP2K, JPEG, WN, Blur, FF
Waterloo IVC 3D P-I [117]	6	1080 x 1920	YES	YES/YES	YES	JPEG, WN, Blur
Waterloo IVC 3D P-II [116]	10	1080 x 1920	YES	YES/YES	NO	JPEG, WN, Blur
MCL-3D [106]	9	1920 x 1080 / 1024 x 728	YES	YES/YES	YES	JP2K, JPEG, WN, Blur, DB, TR, RE
NBU 3D I [95]	N.A	N.A	NO	YES/YES	NO	JP2K, JPEG, WN, Blur, FF
MICT 3D IQA [126]	N.A	N.A	NO	YES/NO	NO	JPEG
SVBL 3D IQA [118]	N.A	N.A	NO	NO/YES	NO	WN, JPEG, JP2K

1.9 Objective quality assessment

Subjective experiments can convincingly assess image quality but they are usually costly, time-consuming, and thus unsuitable for real-time application. This led researchers to consider proposing alternative methods of measurement that could be quantitative metrics. Where the estimation of perceptual quality is performed automatically on computers using algorithms. This objective assessment concept takes digital contents as input and performs quantitative computation in order to give quality ratings. Automatic evaluation (objective evaluation) metrics offer many advantages, such as rapid assessment, low cost. They are easy to incorporate into image processing systems/applications. For these reasons, a significant amount of research has been devoted to the development of objective evaluation metrics. However, objective 2D and 3D stereoscopic IQA methods can be divided into three different groups, as follows:

- Full-Reference (FR) IQA: This group of methods utilize the reference signal, Such metrics compare the pristine stereoscopic image with its distorted version. The main disadvantage of this group is that in practical applications the reference is often not accessible.

- Reduced-Reference (RR) IQA: This type of methods use only partial information of the reference stereoscopic image. A compromise between the two groups can be found by using the reduced reference (RR) metrics, which integrate certain features extracted from the reference signal for comparison.

- No-Reference (NR) IQA: Also known as blind methods, it is the most difficult for researchers to design, since this group considers the reference stereoscopic image to be completely unavailable.

Comparing these groups, the optimal solution in practice will be the NR metrics that can be deployed for any application. Therefore, this thesis focuses on the NR-SIQA, as the initial stereoscopic images are not present in most realistic circumstances.

1.9.1 Performance evaluation indexes

All quality metrics aim at close approximation of the quality as perceived by the user. Therefore, any proposed metric quality ratings are verified with human opinion scores using comparative indexes. The Video Quality Experts Group (VQEG) [17] ITU [1, 3] have provided guidelines of evaluation procedures and shares criteria to evaluate performance of metrics. A critical aspect from ITU recommendation for performance evaluation that describes the importance of mapping the predicted scores by a metric to a common scale with the MOS scores obtained from the subjective experiment. The recommendation allows a simple linear mapping as well as other monotonic mapping procedures such as third order polynomial mapping or logistic mapping. However, the logistic function [100] is the most common mapping used by researchers in IQA field. This function is based on five parameters (θ_1, θ_2, θ_3, θ_4 and θ_5). The used logistic mapping function for the nonlinear regression is introduced by equation 1.4.

$$Q_{map} = \theta_1 \left(\frac{1}{2} - \frac{1}{\exp\left(\theta_2 \left(Q - \theta_3\right)\right)} \right) + \theta_4 Q + \theta_5 \tag{1.4}$$

Where Q and Q_{map} are the objective quality scores before and after the nonlinear mapping, and θ_i ($i = 1$ to 5) are selected for the most excellent fit.

In the following, We define the common performance indexes for IQA metrics:

- **Root Mean Squared Error (*RMSE*):** It is the simplest indicator used to measure metrics' accuracy. The calculation is defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Q_{obj} - Q_{sub})^2} \tag{1.5}$$

  With N refers to the number of images, Q_{obj} and Q_{sub} are objective and subjective scores, respectively. The higher RMSE value corresponds to worse accuracy.

- **Pearson Linear Correlation Coefficient (*PLCC*):** The similarity between data sets is an indicator of how well they relate to each other. The most famous indicator of correlation in statistics is the Pearson Correlation. The PLCC shows the linear relationship between two sets of data, where a PLCC = 1 (usually +1) represents absolute correlation and PLCC = 0 for totally uncorrelated series. This indicator is defined as follows:

$$PLCC = \frac{\sum_{i=1}^{N} (Q_{sub}(i) - \overline{Q_{sub}}) \times (Q_{map}(i) - \overline{Q_{map}})}{\sqrt{\sum_{i=1}^{N} (Q_{sub}(i) - \overline{Q_{sub}})^2} \times \sqrt{\sum_{i=1}^{N} (Q_{map}(i) - \overline{Q_{map}})^2}} \tag{1.6}$$

  where N is the total number of stimuli in the set, Q_{sub} is the subjective scores and Q_{map} is the mapped score obtained using the function 1.4. $\overline{Q_{sub}}$ and $\overline{Q_{map}}$ are the correspond mean values of each set.

- **Spearman's Rank-Order Correlation Coefficient (*SROCC*):** It is the non-parametric version of the Pearson correlation coefficient. It is typically used either for ordinal variables or for continuous data. SROCC = 1 refers to perfect correlation between sets, while SROCC = 0 indicates no correlation. The formula of SROCC is fiven as follows:

$$SROCC = 1 - \frac{6 \times \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{1.7}$$

  where d_i is the difference between the rank of the i-th stimulus in subjective and objective evaluation. For example, if the i-th stimulus has the third highest Q_{sub} but fifth highest Q_{obj}, $d_i = 5 - 3 = 2$.

- **Kendall's Rank Order Correlation Coefficient (*KROCC*):** Another measure used to measure the ordinal association between two measured quantities. To compute KROCC, the order of each pair of stimuli in the set N after both subjective and objective evaluation is checked. If the order in terms of Q_{sub} and Q_{obj} agrees, the pair is considered concordant. In the opposite case, the pair is discordant. The following formula is used:

$$KROCC = \frac{N_c - N_d}{\frac{N}{2}(N^2 - 1)} \tag{1.8}$$

  where N_c and N_d are the numbers of concordant and discordant pairs in the set, respectively.

Overall, the PLCC and RMSE assess the metric prediction accuracy while *SROCC* evaluates the prediction notability degree. Higher values for PLCC, SROCC (close to 1) and lower values for RMSE (close to 0) indicate superior linear rank-order correlation and better precision with respect to human quality judgments, respectively. For a perfect match between the objective and subjective scores, PLCC = SROCC = KROCC = 1 and RMSE = 0.

More recently, a statistic indicator called T-test is being utilized for statistical performance comparison between the IQA metrics [91]. It questions whether the difference between the groups represents a true difference in the study or if it is likely a meaningless statistical difference, where 1 indicates that the groups are statistically different and 0 indicates that the groups are statistically similar.

The T-test analysis is based on calculating the Fisher z-transform value of the correlation coefficient (e.i PLCC, SROCC) as:

$$F_z = \frac{1}{2}\ln(\frac{1+C}{1-C}) = \operatorname{arctan}(C) \tag{1.9}$$

With C is the correlation coefficient. For instance, when comparing two PLCC values with their Fisher z-transform values F_{z1} and F_{z2}, the hypothesis testing approach is employed in order to determine the significance of the difference. Hypothesis H_0 assumes that the two coefficients are not different. The alternative hypothesis H_1 assumes that there is a significant difference between the PLCC values but does not discriminate which one is better. Then, the T value is calculated as follows:

$$T = \frac{F_{z1} - F_{z2} - \sigma_{(F_{z1}-F_{z2})}}{\mu_{(F_{z1}-F_{z2})}} \tag{1.10}$$

Since H_0 assumes no difference, $\mu_{(F_{z1}-F_{z2})} = 0$ and $\sigma_{(F_{z1}-F_{z2})} = \sqrt{\sigma_{z1}^2 + \sigma_{z2}^2}$. The T value is then compared to the 95% t-Student value for two-tailed test with N degrees of freedom. If it is larger, the H_0 can be rejected since the statistically significant difference between the PLCC values has been found. In the opposite case, the hypothesis cannot be rejected.

1.9.2 Objective IQA metrics applications

The 2D/Stereoscopic IQA metrcis can be used to evaluate/optimize the efficiency of 2D/3D processing algorithms/systems (e.g., compression, enhancement). The aim of IQA is to measure automatically the perceived content quality, which is likely to be degraded in different ways. Therefore, a valid IQA approach can evaluate the perceived quality that is highly associated with human quality assessments like the DMOS/MOS.

In many image processing systems/algorithms, there are certain parameters that need to be determined by users to yield the best results. This is often a difficult task for naive users as the best values may be image dependent. A good 2D/stereo IQA measure could be a useful tool to help decide on these parameters automatically. This is illustrated in Fig. 1.16, where depending on the application, either NR, RR, or FR 2D/stereo IQA measures could be employed to create the feedback control signal. For instance, in the case of image enhancement, the NR approach may be used and only the image obtained at the output end is allowed for quality computation. In image coding applications such watermarking/compression, a FR 2D/stereo IQA approach could be used that requires both decoded image from the output end and the original reference 2D/stereo image from the input (referred by the dashed line).

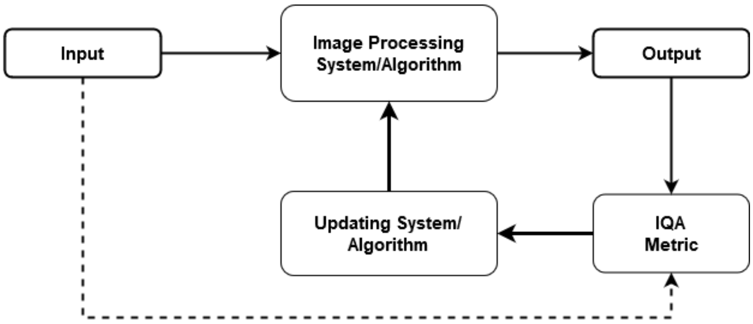


Figure 1.16: Diagram of IQA metric based feedback-optimization.

A special feature of many 2D/stereo IQA metrics that are often ignored by researchers is that they not only have quality ratings, but also produce quality maps that show local quality differences across the picture space. These quality maps can help to identify where in the image the enhancement yield the most improvement.

1.10 Conclusion

Estimation of the stereoscopic image quality is the key factor in design and optimization of 3D visual content system/algorithm. For that, measuring the quality of such content is crucial. Compared to subjective quality assessment, objective estimation is important nowadays for the advantages that offers and for the continues increasing amount of stereo images.

In this chapter we gave an overview on human visual system as well as digital stereoscopic imaging system. Afterward, we identified and described the most common artefacts which could affect a 3D stereoscopic contents. We have also shown that the effect of these distortions on 2D images and stereo-pair images is not the same due to the concealed depth information in the stereoscopic image. In the next chapter, we will introduce a state-of-the-art quality evaluation metrics designed for stereoscopic images.

Chapter 2

Background and Related work

2.1 Introduction

In stereoscopic 3D multimedia systems, visual quality is the main factor that affects the overall quality of experience (QoE) of users. Therefore, the interest in objective SIQA has been growing at an accelerated pace over the past decade. In order to view the attention of stereoscopic image quality over the last twenty years. On Google scholar search engine, we use two keywords to find the number of articles accessible online. Fig. 2.1 shows the growth of scientific articles that mention stereoscopic image quality assessment.

The latest progress on developing automatic SIQA methods may involves multidisciplinary topic. This new progress in both theoretical development and novel techniques appears to be a converging point from a wide range of research directions: computer vision; machine learning; visual system; neural physiology and so forth. While the field of objective SIQA is still evolving rapidly, a novel and better SIQA metrics will continue to emerge in the coming years.

The goal of this chapter is to introduce the different metrics available for stereoscopic 3D images analysis, and provide a framework that can be used as a starting point for those who are interested in developing their own stereoscopic image quality metrics. In particular, we discuss the basic principles of machine learning techniques deployed in IQA domain. Then, we address the state-of-the art SIQA metrics. It is important to mention that this chapter does not provide a complete overview of all the available stereoscopic image quality assessment techniques, but instead focuses on the state-of-the-art techniques. It also provides common guidelines on how image quality assessment metrics can be made.

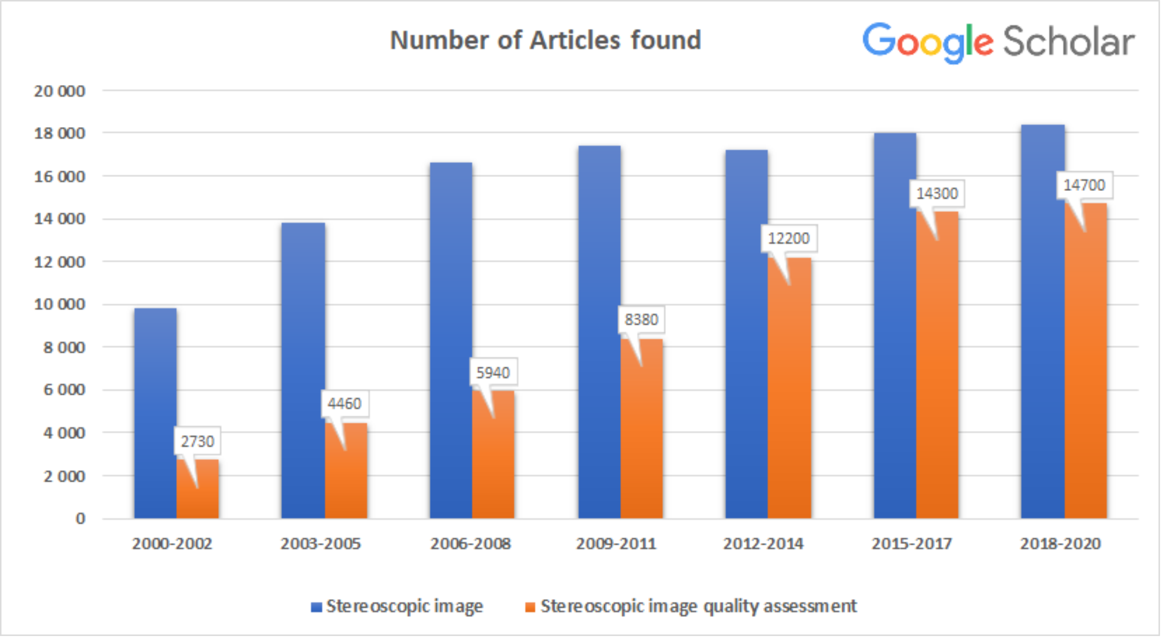


Figure 2.1: Search results at specific periods using Google scholar for the keywords: stereoscopic image, stereoscopic image quality assessment.

2.2 Principles of Machine learning algorithms

With the rise of machine learning algorithms, many of them have been adopted in IQA field of research to propose new state-of-the-art metrics. Among several machine learning techniques, we discuss the most deployed for 2D/Stereoscopic IQA approaches in the following subsections.

2.2.1 Types of machine learning algorithms

The types of machine learning algorithms are mainly divided into four categories: Supervised learning, Un-supervised learning, Semi-supervised learning, and Reinforcement learning.

- Supervised: In machine learning supervised learning is the most common paradigm. It is the simplest to grasp and the easiest to execute. It is designed to learn by examples. The data in supervised learning algorithm is composed of inputs combined with the right outputs. The algorithm can look for patterns in the data during training and form a correspond model. The goal this model is to predict the correct label for newly presented input data. A supervised learning algorithm can be written in its most simple form, simply as: $y = f(x)$, where x is the input, y the predicted output, and f the complex function of the model usually referred as black box.

- Un-supervised: non-supervised learning algorithms can distinguish patterns in data sets containing data points that are neither classified nor labeled. In other words, allow the machine to self-identify patterns within data sets. The most common form of unsupervised learning is clustering which is the process of organizing objects into groups whose members are in some way identical.

- Semi-supervised: Semi-supervised learning falls between unsupervised learning and supervised learning. It is a special instance of weak supervision. Generally, the learning algorithms combines a small amount of labeled data with a large amount of unlabeled data during training. This method can make predictions more accurate than unsupervised and is the most commonly used method. For instance, if there are 1000 photos, 100 of them which are labeled. Through the characteristics of these 100 photos, the machine identifies and classifies the remaining photos. Because there is already a basis for identification.

- Reinforcement: This type is where the machine uses observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk. Reinforcement Learning focuses specifically on letting models learn from mistakes. It is about learning what to do and how to map circumstances to actions. The end goal is to optimize the importance of the incentive signal.

Among the four types mentioned above, when labeled data is available, supervised learning is often the best solution for both classification and regression problems. Therefore it is the common type for IQA metric designs. In particular for NR metrics, where learning-based regression techniques are deployed (e.g., Support Vector Regression (SVR) [31], Gaussian

Process Regression (GPR) [75], Artificial Neural Network (ANN) [107], and Random Forest (RF) [67]). However, we discuss in the following subsections the most supervised learning techniques that are being deployed for 2D/Stereoscopic IQA approaches.

2.2.2 Support vector machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that mainly used for classification problems [111]. But it can also solve regression tasks such image quality evaluation (e.g. SVR). There are two forms of SVRs, depending on the data inputs for the model: linear or non-linear model. However, the problem of regression is to find a function $f(x)$ that approximates mapping from an input x_i domain to real numbers y_i on the basis of a training sample. Therefore during training, SVR algorithm tries to fit the best line within a predefined threshold error value ε as shown in Fig. 2.2. Where any error is permissible as long as it is less than ε.

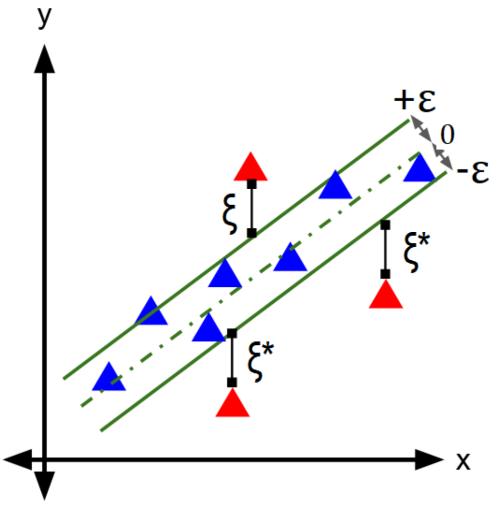


Figure 2.2: Linear SVR example using random dataset.

SVR creates two boundary lines (represented by green line) with the help of best fit line, the hyperplane (represented by doted green line) that has a maximum number of points. Where both the boundary lines are at equal distance (epsilon) from the hyperplane. For the hyperplane $f(x) = W.x + b$, where W refers to weights vector and b is constant to be defined. With C a deviation parameter of the error ε, do the following:

$$MIN : \frac{1}{2} \parallel W \parallel^2 + C \sum_{i=1}^{N} |\xi_i + \xi_i^*| \tag{2.1}$$

$$With \quad constraints : y_i - W.x_i - b \leq \varepsilon \quad and \quad W.x_i + b - y_i \leq \varepsilon$$

For the nonlinear case, the model follows the same steps as in the linear case. However, a change of space for the x_i data is required. A kernel functions are used to transform the data into a higher dimensional feature space to make it possible to perform the linear separation.

2.2.3 Artificial Neural Networks

Artificial Neural Network (ANN) is biologically-inspired network of artificial neurons designed to perform particular functions. It is based on a series of linked units or nodes called neurons, each connection will send a signal to other neurons. These nodes are then used to create layers which together form ANNs [129]. The ANN models have different forms and sizes, but they often include three kind of layers (e.g input, hidden, and output layer). However, ANNs can be deployed for both regression and classification problems, where this form of model has been adopted by several 2D/Stereoscopic IQA metrics.

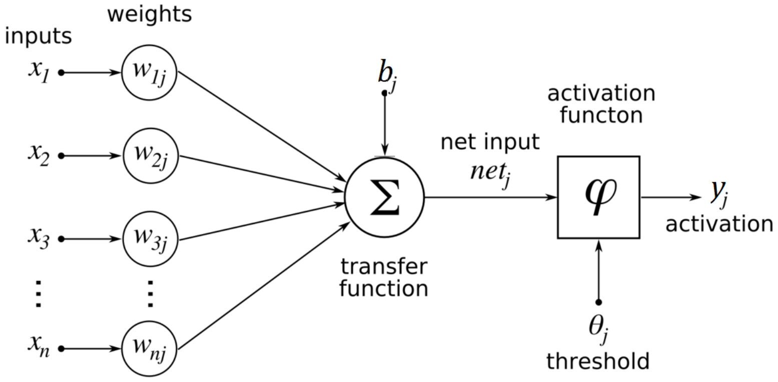


Figure 2.3: Single artificial neuron diagram.

Nearly all artificial neurons can be described by the diagram in Fig. 2.3. For each neuron, the input is multiplied by a weight associated with the connection and then each weighted input is summed before being passed into an activation function. Described by the following:

$$y_j = \varphi(net_j) \quad where : net_j = \sum_{1}^{n}(x_n.w_{nj}) + bj \tag{2.2}$$

Where $X = [x_1...x_n]$ is the input vector. $W_j = [w_{1j}...w_{nj}]$ is a weight vector and b_j refers to a bias value. The activation function φ may changes from model to model. For the simplest activation function, the output y_j of j^{th} neuron is set as one if the sum of the weighted inputs net_j is greater than an internal threshold θ_j or set to zero otherwise.

To use an artificial neural network, it must first be trained for a specific task. The weight vector and bias value are adjusted during training process for each neuron in the network. There are several training approaches for artificial neural networks found in the literature and each has its own advantages and drawbacks for certain tasks. However, the most popular training methods are Back Propagation [48], Restricted Boltzmann Machines [65].

2.2.4 Convolutional Neural Networks

Convolutional neural networks also known as CNNs [6], are widely used for visual imaging quality. They are a specific type of ANNs that are generally composed of the following layers:

- Convolution layer: The convolution layer (CONV) uses filters that perform convolution operations as it is scanning the input image with respect to its dimensions. The parameters of this layer include: the number of filters, size of filters, stride. The resulting output is called feature map or activation map.

- Pooling layer: The pooling layer (POOL) is a down-sampling operation, typically applied after a convolution layer, which does some spatial invariance. After choosing pooling kernel and stride, max or average pooling can be applied where the maximum and average value is taken, respectively.

- Fully connected layer (FC): The fully connected layer (FC) operates on a flattened input where each input is connected to all neurons. If present, FC layers are usually found towards the end of CNN architectures and can be used to optimize objectives such as class scores.

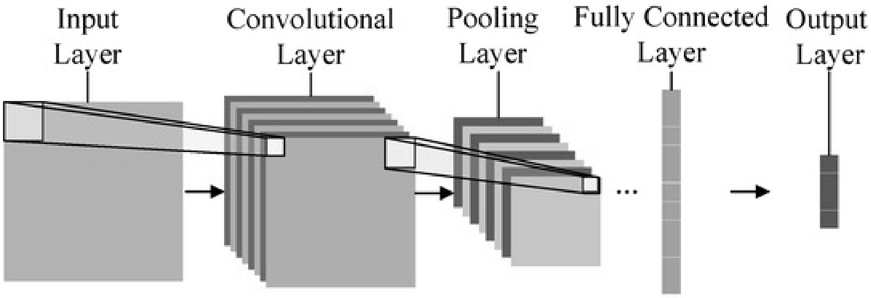


Figure 2.4: Convolution neural network using single layer

An example of CNN model is shown in Fig. 2.4 using each of the layers above. However, CNN architectures mostly include more layers that make the model deeper. With the help of an input and output layers and training algorithm, providing labeled images allow the filters of convolution layer to be learned during the training phase. After the training, the CNN model is fitted on a specific task according to the data. The model is then deployed for prediction on new data rather than the training one.

2.2.5 Convolutional Encoder-Decoder Networks

Encoder-Decoder is a machine learning technique that compresses the input into a feature vector called latent-space representation, and then reconstructs the output from this representation. A convolutional Encoder-Decoder network is a specific type of this technique, where it aims to get latent-space vector from an input 2D image/map using encoder, then using the latent-space vector as input for a decoder, it generates (same or different) image/map. The convolutional Encoder-Decoder is used for various different applications such as image segmentation, disparity map estimation, generative models and so forth. Also, it has been utilized recently in IQA domain. However, as shown in Fig. 2.5 two parts of this kind of network can be simply defined as:

- Encoder: This is the part of the network that compresses the input into a latent-space representation. It can be represented by an encoding function $h = f(x)$.

- Decoder: This part aims to construct an image/map output from the input latent space representation. It can be represented by a decoding function $y = g(h)$.

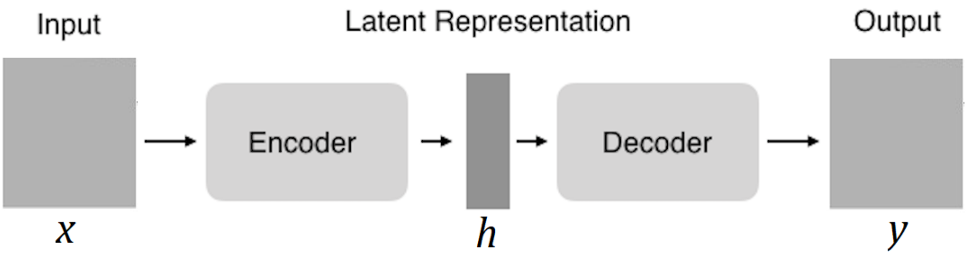


Figure 2.5: Architecture of an Encoder-Decoder network

The model as a whole can thus be described by the function $g(f(x)) = y$. The training process aims to define the weights of this model (e.g encoder and decoder convolution network weights). However, this technique can be used in the IQA domain to extract relative quality features, reduce feature dimensionality, generate distortion maps, and so on [29].

2.3 Concept types of SIQA metrics

Most SIQA approaches, regardless of their core concept, have three major phases as shown in Fig. 2.6. The first phase is to preprocess the stereoscopic image data in order to extract valuable and efficient information as easily as possible. This phase may include filtering, color conversion, domain transform, normalization, scaling and so on. The second and most important phase is the feature extraction/learning process, which involves either manual/handcrafted or automated feature extraction. The handcrafted features are fitted using machine learning-based regression (e.g., SVR, ANN, KNN and so forth) while the automatic features are extracted and controlled by an end-to-end (deep learning) prediction models (e.g. CNN, Encoder-Decoder-CNN, ...etc). There are two choices for quality features: global or local features. Global features describe the scene as a whole (e.g Histogram Oriented Gradients (HOG) [27], contour representations, shape descriptors, and texture features) to the generalize the entire scene while the local features describe a patches from the scene (SIFT [70], SURF [11] features). The final phase is to compute the quality score based on the learned/regression model with or without the use of human opinion ratings MOS or DMOS values (i.e., supervised or unsupervised models).

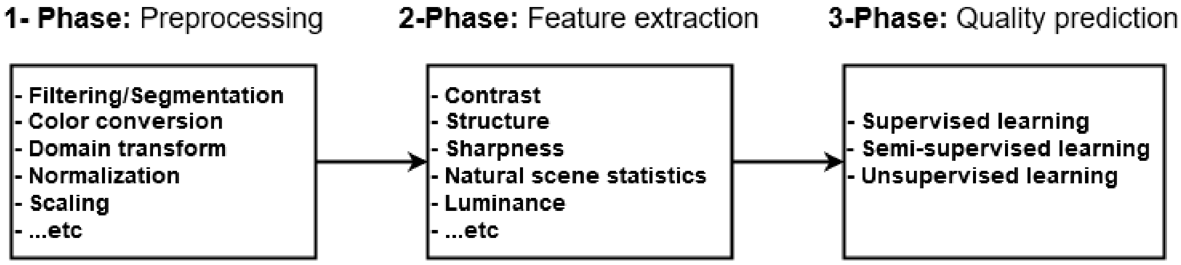


Figure 2.6: The most followed three phases of SIQA metrics.

Several concepts can be followed to design SIQA metric such as HVS models, Natural Scene Statistics (NSS) computation models, and depth based models. The design of recent 2D/Stereoscopic IQA metrics usually incorporate machine learning techniques as discussed in previous section. While some of the proposed 2D/Stereoscopic IQA metrics, mainly older ones did not deploy machine learning algorithms. Therefore, the SIQA designs can be classified into two classes. The first design directly applies the 2D QA models to the SIQA problem by simply calculating the mean quality predicted of left and right views. The second class take disparity/depth information into account while designing the metric. However, many researches support that the quality of 3D contents are not deduced accurately from the average of the two views quality scores [97].

The human observer is the ultimate receiver of stereoscopic contents. Therefore, the research focus is shifting towards developing SIQA methods which exploit knowledge about the HVS rather than only using quality factors such as contrast, luminance, distortions...etc. The use of deep learning techniques also seems to be as a promising direction in the future. It can lead us toward a third class of SIQA metrics where they do not necessarily rely on explicit models but on data-driven approaches that allow end-to-end optimization. For instance, authors in [131] The authors defined the architecture of a CNN model and tuned its parameters for 2D IQA before deploying it for SIQA, where left view, right view, and difference were given as inputs. In the following subsections, we address the state-of-the-art SIQA metrics by their type of concept.

2.3.1 2D IQA metrics

Perceived stereoscopic image quality can be obtained by standard quality metrics for 2D images. These metrics are extended to stereoscopic 3D imaging systems, without using the additional advantage of stereoscopic depth. The earliest and most widely used FR 2D IQA metrics are the mean square error (MSE) and peak-signal-to-noise-ratio (PSNR), which simply quantify the difference between the reference Ir and the distorted Id images, respectively. PSNR and MSE are respectively defined as follows:

$$MSE = \frac{1}{N}\sum_{n=1}^{N} \parallel I_r - I_d \parallel^2, \tag{2.3}$$

$$PSNR = 10 \cdot \log_{10} \cdot \frac{P_{max}^2}{MSE} \tag{2.4}$$

where N denotes the pixels number, and P_{max} is the maximum pixel value of the reference image. Although PSNR is still widely used, it has a poor correlation with the human judgment of quality due to lack of consideration of the HVS properties. A number of objective 2D IQA measures have come after and showed consistent performance that outperforms MSE and PSNR in terms of correlations with subjective quality evaluations. For example, Wang *et al* proposed the Universal Quality Index (UQI) metric that is defined as:

$$\begin{aligned} UQI(I_r, I_d) &= \frac{1}{M}\sum_{n=1}^{M} UQI_{map}(I_r, I_d), \\ &= l(I_r, I_d) \times c(I_r, I_d) \times s(I_r, I_d) \end{aligned} \tag{2.5}$$

where M is the number of local windows with size 8 x 8 .While $l(.)$, $c(.)$, and $s(.)$ refer to luminance, contrast, and structural/correlation similarities between I_r and I_d, respectively, given as follows:

$$\begin{aligned} l(I_r, I_d) &= \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1}, \\ c(I_r, I_d) &= \frac{2\sigma_r\sigma_d + C_2}{\sigma_r^2 + \sigma_d^2 + C_2}, \\ s(I_r, I_d) &= \frac{\sigma_{rd}}{\sigma_r\sigma d} \end{aligned} \tag{2.6}$$

Whereas μ_r and σ_r denote the average and variance of I_r respectively, and σ_{rd} is the covariance of I_r and I_d. While C_1 and C_2 are constants to avoid zero value in the denominator. Wang *et al* [120] followed the idea of UQI [119] and developed a structural similarity (SSIM) index, assuming that the HVS is sensitive to the structural information of a scene. In fact, SSIM highlighted the importance of HVS properties for the design of IQA metrics. Based on this finding, various accurate 2D IQA models have been proposed over the last decade.

Generally, all 2D IQA metrics can be enhanced to stereoscopic images. These 2D-extended metrics usually extract feature vectors separately for the left and right images. They are weight-averaged to obtain the final feature vector for training. In the meanwhile, other improved 2D IQA metrics tend to use the disparity/depth map either by adding it in the feature extraction process or by incorporating it into the original design. For instance, Gorley *et al.* [45] did not use or measure disparity/depth information. They compute quality scores on matched feature points delivered by SIFT (Scale-Invariant Feature Transform) [70] and RANSAC (RANdom Sample Consensus) [40] applied to the left and right views. SIFT and RANSAC are considered the most efficient algorithms for extracting and matching features. However, the overall results of these extended metrics are not as good as for 2D images, which motivates to have metrics dealing with 3D perception.

2.3.2 SIQA metrics based on depth perception

Depth information has direct influence on stereoscopic image quality as discussed in previous chapter (section 1.6). Therefore, several existing SIQA metrics are based on depth map or have incorporated this information in the design of the model. Fig. 2.7 shows the basic workflow of SIQA metrics that adopt disparity/depth information. The dashed line in the figure refers to an additional inputs that the metrics may take.

For example, Akhter *et al.* [5] have designed a no-reference stereo IQA algorithm. Based on the assumption that the visual distortion and disparity of any stereoscopic display is highly reliant on local features, such as edge (non-plane) and non-edge (plane) regions.

They extracted a combination of features from disparity map and stereo image pairs. Benoit *et al.* [12] proposed a FR SIQA from 2D IQA enhanced metrics using disparity information. The proposed metric relies on SSIM and C4 [19] metrics. Where both these 2D IQA methods are based on the comparison between the structural information extracted from the distorted and the original images. The authors then added the Euclidian distance between distorted and original disparity maps for comparison. A similar FR SIQA has been suggested by You *et al.* [130]. They used a variety of 2D IQA models for stereoscopic images and tried to combine the predicted quality scores from both disparity map and stereo pair. The authors of [45, 50] proposed a PSNR-based stereo IQA metrics. While Hewage *et al.* [50] computed the edges from the disparity map, then PSNR has been used between the pristine and test edge maps to predict the quality.

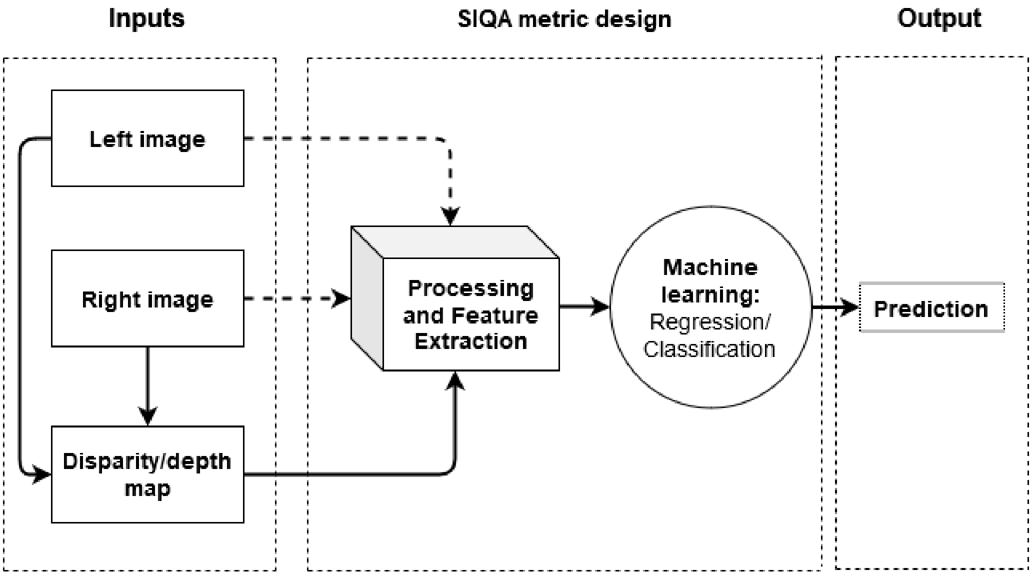


Figure 2.7: The basic workflow of SIQA metrics based on disparity/depth map. The dashed line denotes optional inputs.

Overall, the results obtained of the above-mentioned metrics and their ablation analysis have showed the benefits of the disparity/depth map, where the workflow takes into account the depth perception.

2.3.3 SIQA metrics based on naturalness

Higher level assessment concepts such as naturalness and viewing experience are proposed that are sensitive to both image quality and stereoscopic depth. This high level

3D quality evaluation models are constructed in which the quality is related to viewing experience and naturalness terms. NSS model assumes that natural scenes possess certain regular statistical properties and is widely used in NR SIQA metrics. Fig. 2.8 shows the basic workflow of SIQA metrics that adopt naturalness information. The dashed line in the figure refers to an additional inputs that the metrics may take. Several existing NR SIQA metrics extract NSS-based features and then deploy machine learning technique for quality score prediction. The NSS-based features are obtained from studying the variation of image statistics, which are characterized by the fitting parameters of NSS model, across different distortions. However, distortions not only change the stereoscopic image statistics, but also disturb the statistical regularity held by natural ones. The NSS-based features is therefore closely associated to quality and even can be used to detect the type of distortion. Furthermore, research has shown that the added benefit of stereoscopic depth is integrated substantially more by naturalness.

This 3D image quality evaluation concept is therefore widely adopted and has proven to be reliable in SIQA algorithms. For example, Moorthy *et al* in [83] improved their blind 2D IQA method for stereoscopic images. Their design includes distortion detection, followed by distortion-specific quality evaluation using NSS features. The performance indicated superiority of the method compared to PSNR and statistically equivalent to the popular SSIM. While Su et al. [109] built a NR SIQA framework. They synthesized a HVS viewing model and then they extracted bivariate and generalized univariate NSS as features. More NSS-based framework has been conducted by Appina *et al* [7]. Authors utilized a bivariate generalized Gaussian distribution (BGGD) model to fit the distribution of luminance and disparity coefficients. Then, a fitting parameters have been used as final quality-aware features. Lv *et al* [71] also developed a NR stereo IQA metric. Their scheme computes binocular self-similarity and binocular integration using NSS features. Overall performance results of theses NR SIQA metrics classify naturalness as an appropriate concept to evaluate the quality of stereoscopic images.

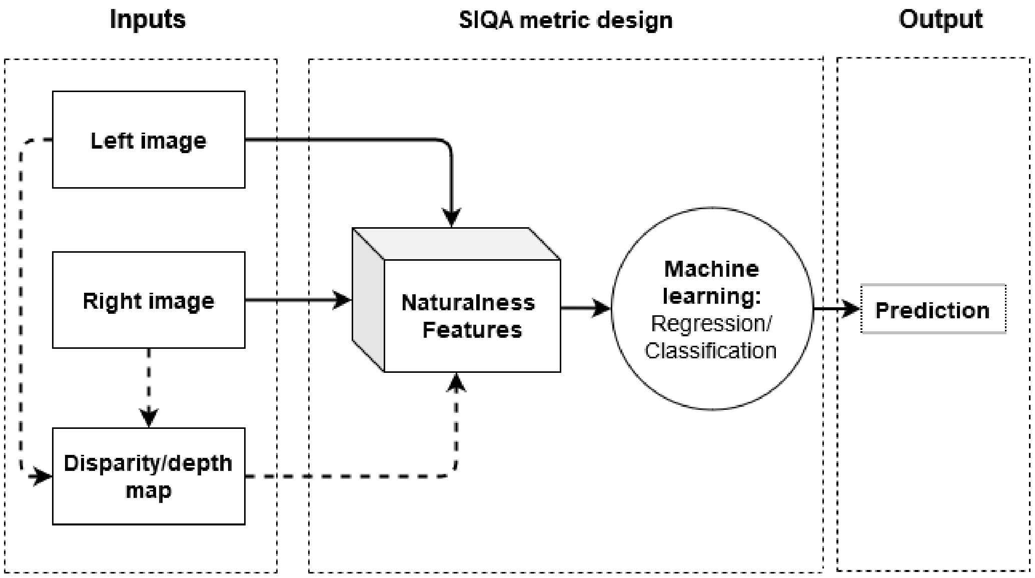


Figure 2.8: The basic workflow of SIQA metrics based on NSS quality feature. The dashed line denotes optional inputs.

2.3.4 SIQA metrics based on Human visual system modeling

The accurate measurement of visual quality as it is perceived by humans is crucial for any visual communication or computing system in which humans are the ultimate receivers. Traditional HVS approaches to visual quality estimation were based on characteristics that human take the most for judgment such as scene structure, sharpness, luminane, contrast and so forth. These former models consider HVS as black box and tend to optimize quality features related to HVS. While the recent HVS based SIQA models try to simulate or mimic the binocular vision processing of human, they translate the HVS to mathematical model that takes left and right view as inputs. Fig. 2.9 shows the basic workflow of SIQA metrics that adopt HVS modeling. The dashed line in the figure refers to an additional inputs that the metrics may take. However, the HVS is a complex visual process and still an open question for researchers. For this challenging problem, many researchers have used fusion hypothesizes of the perceived left and right eye signals called cyclopean view.

A cyclopean image/view is a single mental representation of a scene generated by the brain after integrating two images obtained by both eyes. The conceptual mechanism that generates the cyclopean image is critical for stereo vision. It is logical way to solve the problem of HVS simulation for SIQA designs. Where, hypothetically the quality estimation is done over the merged single view created in the human brain (e.g the cyclopean image). The majority of the cyclopean image used in SIQA designs are constructed based on endowing weights inspired from Levelt et *.al*[66]. In fact, Levelt et *.al* have avoided computing cyclopean image. But they were the first researchers to come up with the idea of assigning weights to left and right views of stereoscopic scenes in order to account for binocular rivalry. Where the model is given as follows:

$$Q = W_L Q_L + W_R Q_R \tag{2.7}$$

where Q_L and Q_R are the quality of the left and right views, and Q is the quality score of the stereoscopic image. The weights are usually normalized, it is ensured that:

$$W_L + W_R = 1 \tag{2.8}$$

This simple model, however, is still in use because it shows not only simplicity, but also the possibilities that can be taken. For example, several SIQA metrics have replaced the quality scores Q_L and Q_R with left I_L and right I_R, respectively, to construct a cyclopean image. While the weights also depends on the ideas of algorithm designers, but most of them are based on filtering (e.g., Gaussian, Gabor, Laplacian of Gaussian,...etc), local/global energy estimation, information entropy models, and so on. In which these weights are accomplished under the guidance of binocular fusion that considers binocular rivalry phenomenon.

A significant amount of articles have been published into how the visual system receives the signals perceived by the two eyes, and these studies are currently being used to solve the SIQA problem. For instance, a novel FR SIQA metric called binocular energy quality metric (BEQM) has been proposed by Bensalma *et al* [13]. It estimates the quality by computing the binocular energy difference between the original and distorted stereopairs taking into account the fusion process of the human perception. The basic idea is to construct a model that can replicate the binocular signal produced by simple and complex cells, as well as estimate the related binocular energy. However, the method has shown high correlation with the human judgement. Chen *et al.* [21] proposed a FR quality assessment model that utilized the linear expression of cyclopean view [66] influenced by binocular suppression/rivalry between left and right views. An extended version of this framework has been used to create a NR model using natural scene statistics features extracted from stereoscopic image pairs [22].

In [24], the author has also used the cyclopean image hypothesis and proposed an FR SIQA metric using a 2D FR-IQA fusion. In [46], another FR metric has adopted HVS modeling for stereoscopic images. Where they used Binocular Just Noticeable Difference (BJND) [133] approach to model the binocular rivalry theory. Fang *et al.* [35]. proposed an unsupervised blind model for stereoscopic images. From the monocular and cyclopean view patches, they extracted various quality-aware features in spatial and frequency domains. Then, Bhattacharyya-like distance has been used to produce a quality score. Furthermore, another referenceless SIQA method proposed in [135] that simulated the main functional structure of binocular vision. Then, a dictionary learning based on log-Gabor filter is used to extract features and k-nearest-neighbors (KNN) has been deployed to map the quality score.

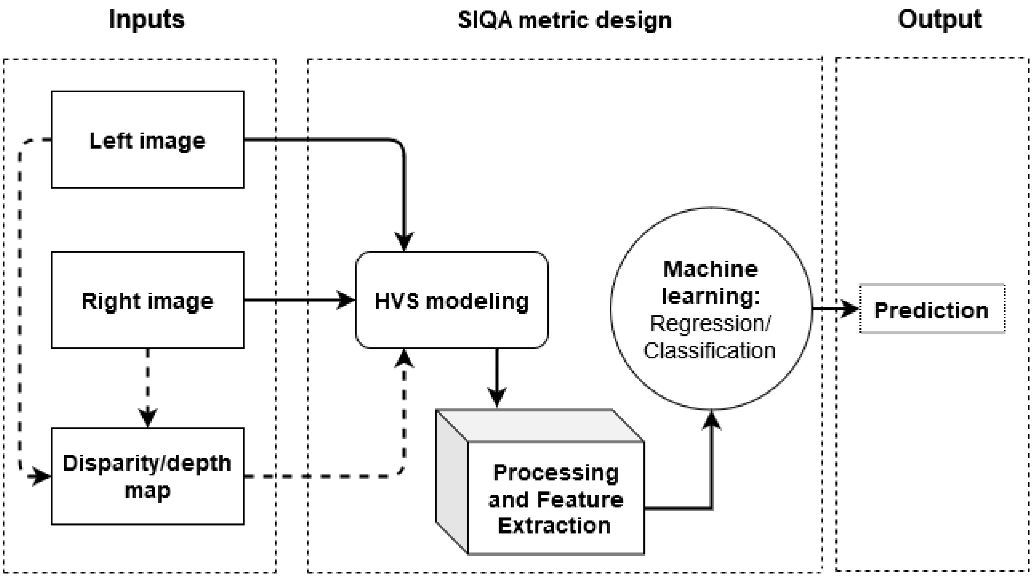


Figure 2.9: The basic workflow of SIQA metrics based on HVS modeling. The dashed line denotes optional inputs.

2.4 Recent SIQA metrics: State-of-the-art

Nowadays, researchers who work on stereoscopic IQA are increasingly relying on NR metrics due to the advantages they offer. It is worth noting that, regardless of their concept, the previously mentioned SIQA methods differ in their approach of extracting quality-aware features, but they all use learning algorithms to create a nonlinear mapping from automated or handcrafted quality features to subjective quality scores. The learning mechanisms also may differ from metric to another, but the success is heavily reliant on the extracted quality features. Another success factor is simulating the quality assessment behavior of the HVS during binocular vision. Whereas the latter is still in its early stages. However, in the following, we briefly address the recent suggested FR, RR, and NR SIQA metrics:

- **FR-SIQA metrics**: A full-reference metric based on binocular receptive field properties has been proposed in [99]. During the training process, the scheme tends to learn a multi-scale dictionary from the training database. In the quality estimation phase, they calculate a sparse feature similarity index based on the estimated sparse coefficient vectors. This latter (e.g coefficient vectors) is built with phase and amplitude differences in mind, as well as a global luminance similarity index that takes luminance changes into account. A similar FR SIQA method was proposed in [72]. This metric is also inspired by human binocular perception, where the binocular perceptual properties of simple and complex cells are simulated. For simple cells simulation, which is assumed to represent a monocular cue, the authors have used a push–pull combination of receptive fields response. While complex cells, which are used to represent a binocular cue, are simulated by using binocular energy response and binocular rivalry response. Following the simulation phase, quality-aware characteristics are extracted from the responses using a self-weighted histogram, and similarity measurement is used to determine the quality score. Furthermore, another recent metric based on monocular and binocular visual features is presented in [102]. First, the authors suggested a segmentation strategy to find occluded and non-occluded areas in the scene by using disparity information and Euclidean distance between stereo pairs. The occluded regions are considered to represent the monocular vision while non-occluded regions to reveal the binocular vision of the HVS. Global and local features are then extracted from the regions and used to predict the visual quality.

- **RR-SIQA metrics**: A metric is presented in [89] by using binocular perceptual information. This latter is represented by the distribution statistics of visual primitives in left and right images, which are extracted by sparse coding and representation. Authors in [74] have characterized the statistical properties of stereoscopic images in the reorganized Discrete Cosine Transform (RDCT) domain to perform an RR-SIQA. In [73], an RR-SIQA method based on NSS and structural degradation has been also proposed.

- **NR-SIQA metrics**: Researchers are becoming more interested in reference-less/blind SIQA metrics. In [25], the authors proposed a new NR SIQA framework based on a degradation identification and fusion steps of features. A similar NR-SIQA was proposed in [38] where the metric scheme first classifies the distortion type before measuring the quality, including symmetrically or asymmetrically distortion cases.

In [68], the authors have explored Singular value decomposition (SVD) computation tool for NR-SIQA metric. Whereas the findings demonstrated the impact of different distortions to the scene's structure, which are expressed by variations in singular values. They first use SVD on the left and right views to obtain singular values and singular vectors, and then extract energy and structure distributions as quality features from the singular values and singular vectors. More recently, an advanced NSS-based features and complex combination were used to develop modern NR SIQA metrics. For example, Karimi *et al.* [57] combined statistical features derived from a synthesized phase/shift and contrast images.

While Deep convolutional network predictors also have being used. For example in [87], a local patches are extracted and then combined to obtain global features using an aggregation layer in the network. In [128], the authors have considered the deep perception map and binocular weight model to predict the perceived stereo image quality. Meanwhile, Zhou *et al.* have suggested a NR-SIQA metric called StereoQA-Net [134] using a novel end-to-end dual convolutional network. Another recent NR-SIQA metrics have been proposed that utilize deep learning technique. For instance, authors in [101] have used deep sub-networks in a single model to extract primary, local, and global features from the input left and right image. These features are eventually concatenated for quality score regression. Another end-to-end deep learning based NR-SIQA metric proposed in [85], the authors have used Siamese architecture, where the model consists of two CNN models in parallel to each other, which have the same structure and share the weights. As previously stated, the encoder-decoder technique could also be used for SIQA metrics. For instance, authors in [127] have modeled the human visual cortex using the deep auto-encoder. Xu *et al.* [124] have simulated our human brain cognition process to propose NR-SIQA metric using the deep encoder-decoder network. Meanwhile in [56], the authors have optimized feature evolution and nonlinear feature mapping by using encoder-decoder model, where NSS-based features were extracted from a synthesised cyclopean image, left and right views.

Meanwhile, authors in [110], deployed saliency information to determine salient and non-salient patches for local features extraction. The authors used reference stereoscopic images to compute local quality maps. These maps are then used as labels to train deep networks. Another work that takes onto account this type of information in [20], where visual saliency, local magnitude, and local phase are extracted from the stereo image as basic feature vectors, which is then utilized for learning the quality assessment. Image segmentation technique also has been deployed for NR-SIQA methods. Where in [69] a superpixel segmentation is used based on K-mean clustering approach [4]. Then, from these superpixel regions, a spatial entropy and NSS features are extracted to obtain quality ratings using regression model.

In the literature, the popularity of NR-SIQA metrics is growing in comparison to the number of FR and RR SIQA metrics, we also notice that only a few RR-SIQA metrics have been proposed. The performance of some of the above metrics is inconsistent with asymmetric distortions. Whereas the design of these metrics does not consider binocular rivalry/suppression nor HVS modeling. However, the current scope of SIQA metrics is toward HVS design concept as well as using deep learning techniques. Where this ML techniques known for its ease of use and provides promising results, while the metrics that adopt HVS modeling perform better on asymmetric distortion. Despite the fact that many deep learning based SIQA models have achieved outstanding performance on particular SIQA datasets, there are several limitations and problems with real-world implementations, such as: involves heavy computation, is sensitive to pixel attacks, has fixed parameters (e.g image input size), and so on. In addition, in most of the suggested SIQA approaches the human visual attention (e.g saliency information) is not explored. The NR-SIQA is still in its early development phase. In the following chapters, we aim to suggest NR SIQA metrics that address the stated drawbacks.

2.5 Conclusion

Since stereoscopic artifacts yield not only visually unpleasant results, but also visual fatigue or pain, the growing amount of stereoscopic 3D content calls for the development of novel reliable stereoscopic IQA metrics to assure good experience for the users. The majority of SIQA metrics in the literature have been developed using machine learning techniques (e.i SVR, CNN ...etc) since they provide good outcomes and ease of use. However, an ideal quality metric should have the following properties:

- 1) Perceptual : mimics the HVS perceptual mechanism.
- 2) Objective accuracy : provide a numerical representation of the quality as perceived by the observers.
- 3) Reliability : provide perceptual quality prediction for wide variety of content, as perceived by a large amount of observers.

Chapter 3

Contributions based on handcrafted quality feature extraction and HVS modeling

3.1 Introduction

The 2D IQA has progressed significantly in recent years, while stereoscopic IQA is still in its early stage. One of the main aspects in stereoscopic images is assuring a good 3D viewing experience for users. Therefore, an accurate and dependable IQA metrics for stereoscopic content must be created. To accomplish this, we present two approaches for stereoscopic images in this chapter, based on handcrafted quality features and HVS modeling. The first method is for measuring quality [78, 81], while the second is for recognizing distortion types [80].

3.2 1st approach: NR-SIQA based on AdaBoost neural network and cyclopean view

3.2.1 Approach overview

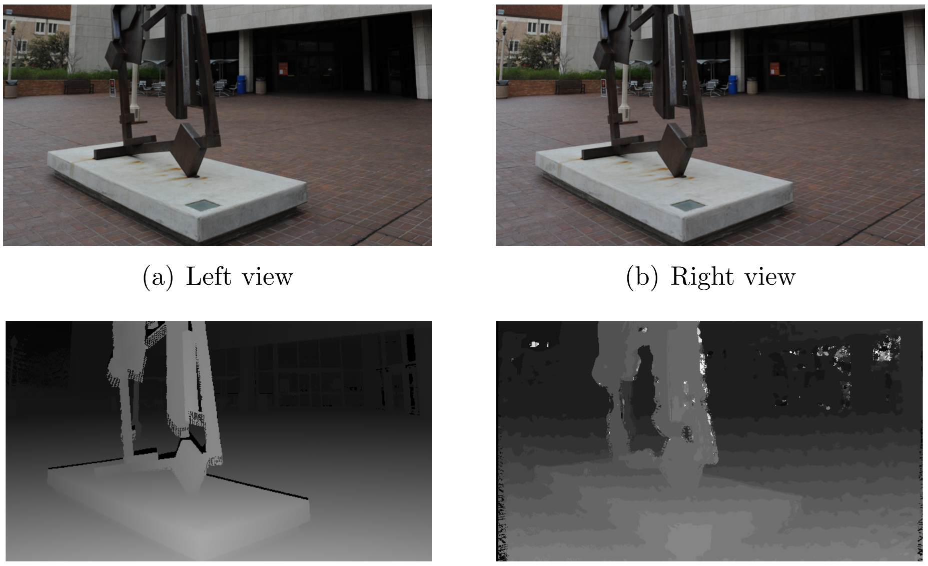
In order to design a model that can assess the quality of stereoscopic images, research on human binocular perception is required. The hypothesis of cyclopean image is therefore used with consideration of binocular suppression. Metrics that use this hypothesis, such as the FR stereo IQA model in [21], have achieved good performance.

The artificial neural network models are widely used for regression and classification [61] along with Back-Propagation (BP) algorithm for training. The proposed model includes an Adaptive Boosting (AdaBoost) technique using with ANN as learners. This techniques has showed robustness and good generalization performances in various applications. Motivated by these ideas we develop a new NR quality predictor model for stereoscopic images. In summary, the model involves three steps: first, a cyclopean image is constructed using Gabor filter responses and disparity map. In a second step, gradient characteristics of the cyclopean image and the disparity map are extracted. Finally, to predict a quality score based on feature learning, the AdaBoost algorithm combined to artificial neural network has been used.

3.2.2 Disparity map computation

The disparity information has proven to be a strong effective factor for stereo images and videos quality. Therefore, it is a necessary information for assessing the quality of the stereo content. Intensive research has been conducted on the design of stereo matching algorithms (disparity estimation). However, there is no agreement, on the type of stereo matching algorithm to be used in stereo IQA, except for those with low complexity. Therefore, a stereo matching model with balanced complexity and performance is deployed.

The chosen algorithm is called SSIM-based stereo. It is an improved version of Sum of Absolute Differences SAD stereo matching algorithm [84]. The modification consists in replacing SAD by SSIM in computing disparities. SSIM [120] scores are used to select the best matches. This is done by maximizing the SSIM scores between the current block from left image and right image blocks along the horizontal direction. The maximum number of pixels to be searched for is the maximum disparity. After all, the disparity map values are the difference between the current pixel and the best SSIM location. A 7 by 7 block size has been used, while the maximum disparity distance has been set to 25. Fig. 3.1 shows an estimated disparity versus the ground-truth disparity using the SSIM-based stereo matching algorithm.



(a) Left view

(b) Right view

(c) Truth disparity

(d) Estimated disparity

Figure 3.1: Top: left and right views of the stereo image. Bottom: Estimated disparity versus the ground truth disparity.

3.2.3 Gabor filter responses

Various theories have been proposed to explain binocular rivalry. This visual phenomenon has recently been investigated by many researchers. Binocular rivalry or suppression is known as failure of the brain in fusing the left and right views causing fatigue or discomfort to the viewers.

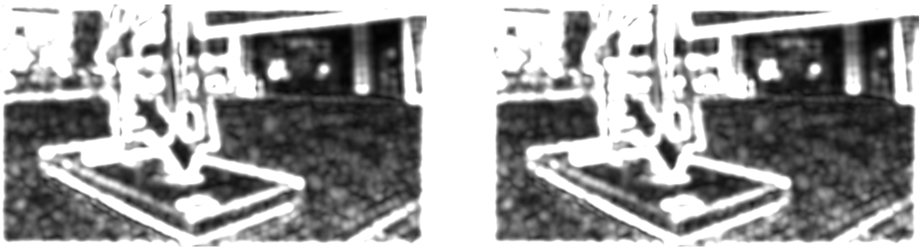
The binocular rivalry as mentioned before (first chapter) is when the two images of a stereo pair present different kinds or degrees of distortion. Therefore, the objective quality of the mostly viewed stereo image cannot be predicted from the average quality of the left and right views. Levelt *et al* have conducted a series of experiments which clearly demonstrate that binocular suppression or rivalry is strongly governed by low-level sensory factors. They concluded that visual stimuli which have more contours or high contrast, tend to dominate the rivalry. Motivated by this result, the energy of Gabor filter bank responses on the left and right images is therefore used to simulate suppression selection (binocular rivalry) of the cyclopean image when it is computed.

The Gabor filter bank is a band-pass filter. It extracts luminance and chromatic channels features. The filter is related to the function of primary visual cortex cells in primates [28]. It models the frequency-oriented decomposition in primary visual cortex, and captures energy in both space and frequency in a high localized way [39].
The used Gabor filter is as follows:

$$\begin{aligned}
&GF(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}[(x'/\sigma_x)^2+(y'/\sigma_y)^2]} e^{i(x\zeta_x+y\zeta_y)}\\
&with\\
&x' = (x-m_x).cos(\theta) + (y-m_y).sin(\theta)\\
&y' = -(x-m_x).sin(\theta) + (y-m_y).cos(\theta)
\end{aligned} \tag{3.1}$$

where m_x and m_y define center of the Gabor receptive field (m_x and m_y are the x and y locations of the center with respect to the original coordinate system). σ_x and σ_y are the standard deviations of an elliptical Gaussian envelope along x' and y' directions, ζ_x and ζ_y are spatial frequencies, and θ orients the filter. The design of the Gabor filter bank is based on the work conducted by Chun *et al.* [108].
In visual perception study, a spatial frequency is expressed as the number of cycles per degree of visual angle. However, in theory, the spatial frequency is that the visual cortex operates not only on the lines and straight edges code but also on a spatial frequency code. To support this theory, a series of experiments have been conducted by P. Issa *et*

al. [55]. They studied the effect of spatial frequency on primary visual cortex reaction using cats. The authors concluded that the visual cortex neurons react even more robustly to sine-wave gratings in their receptive fields at specific angles than they do to edges or bars. Therefore, using a band-pass filter over multiple orientations is favorable to extract features which the visual cortex responds to. The choice of the spatial center frequency is inspired by the result of Schor *et al.* [96] who found that the stereoscopic acuity of human vision normally falls off quickly when seeing stimuli dominated by spatial frequencies lower than 2.4 cycles/degree. Based on their findings, this means that using filters having spatial center frequencies in the range from 2.4 to 4 cycles/degree should produce responses to which a human observer would be more sensitive. Therefore, the local energy is estimated by summing Gabor filter magnitude responses over eight orientations at a spatial frequency of 3.67 cycles/degree ($\zeta_x = \zeta_y = 3.67$). The standard deviations σ_x and σ_y are set to 0.01 ($\sigma_x = \sigma_y = 0.01$). As an example, Fig. 3.2 shows the filter outputs on the left and right views.



(a) Left view

(b) Right view

Figure 3.2: Gabor filter responses from the left and right views.

3.2.4 Cyclopean image construction

The cyclopean image synthesis differs from the usual 2D images for the depth information it contains. The first objective of the proposed stereo IQA algorithm is to estimate the actual cyclopean view formed within the observer's mind while a stereo image is supplied. The HVS processes and combines visual signals from both eyes into a single combined perception [15]. It is worth noting that the HVS has not been completely understood. Therefore, the cyclopean image that is actually processed in our minds is still unclear.

The visual signals from the two eyes are added by HVS, a process called binocular summation which enhances vision and increases the ability to detect weak objects [15]. However, current knowledge of HVS is very modest to guide the development of a mathematical model that perfectly simulates the process in the human brain. Therefore, a popular choice is to replace the complex simulation by simplified mathematical models. In our study, we use a linear model of a cyclopean image has been in order to consider the phenomenon of binocular rivalry.

The model is as follows:

$$C = w_l I_l + w_r I_r \tag{3.2}$$

where I_l and I_r are respectively the left and right images, both w_l and w_r are the weighting coefficients for the left and right eyes in which $w_l + w_r = 1$.

The energy of Gabor filter bank responses is used to compute the weights, while the SSIM-based stereo matching algorithm is employed to create the disparity map.

The used model is:

$$C(x, y) = w_l(x, y) \times I_l(x, y) + w_r(x + m, y) \times I_r(x + m, y) \tag{3.3}$$

where the weights w_l and w_r are given by:

$$w_l(x, y) = \frac{GI_l(x, y)}{GI_l(x, y) + GI_r(x + m, y)} \tag{3.4}$$

$$w_r(x + m, y) = \frac{GI_r(x + m, y)}{GI_l(x, y) + GI_r(x + m, y)} \tag{3.5}$$

where GI_l and GI_r are the summation of Gabor filter magnitude responses from left and right views respectively, and m is the disparity index that corresponds to pixels from left image I_l to those in right image I_r. The filter of the form (3.1) is used to compute the magnitude responses over eight orientations for better accuracy. In the equation (3.1), θ refers to the filter's orientation degree. Table 3.1 shows the used orientation degrees.

Table 3.1: Magnitude responses orientation degrees.

Orientation number	1	2	3	4	5	6	7	8
Orientation degree	0	22.5	45	67.5	90	112.5	135	157.5

Figure 3.3 summarizes the construction steps of the cyclopean image. While Fig. 3.4 shows an example of cyclopean image. In this example, a stereo image without distortion has been used. (Figure 3.1, left view (a) and right view (b)).

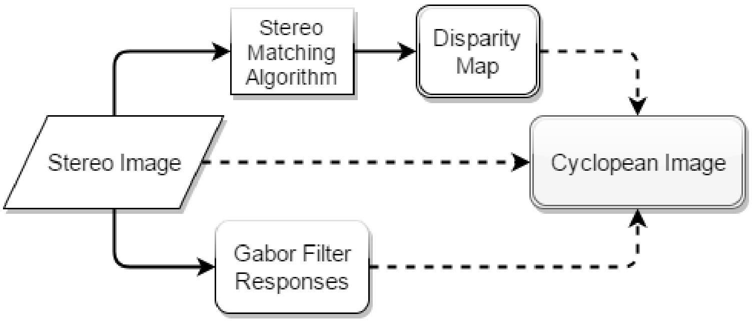
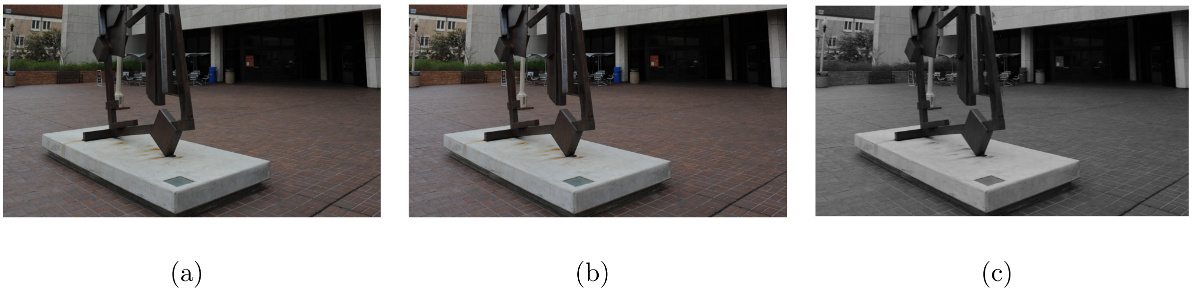


Figure 3.3: The flowchart of the formed cyclopean image.

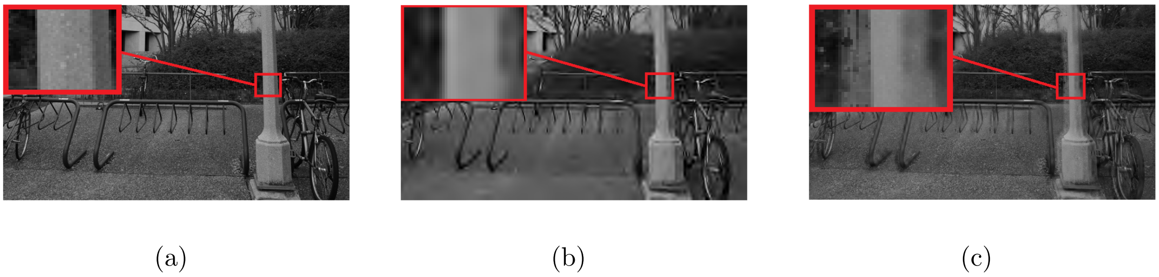

(a)

(b)

(c)

Figure 3.4: (a) Left view, (b) Right view, and (c) the synthesized cyclopean image by the proposed framework.

Another example of cyclopean image obtained from asymmetric distorted stereoscopic view in Fig. 3.5. The outcome cyclopean image computed from the undistorted left image and right image that is distorted. The red boxes in figure zoom into the same location of each view. It can be noted from the figure that the asymmetric distortion is clearly stated in the formed cyclopean image.



(a)

(b)

(c)

Figure 3.5: (a) Left image without distortion, (b) Right image JPEG distortion, (c) cyclopean image of both images. For each view, red box is zoomed to the left for better visualization.

3.2.5 Feature extraction

In addition to screen height and number of displayed pixels, the viewing conditions, namely: visual angle and viewing distance also influence the stereo image quality for the observer. However, the visual angle and viewing range are not taken into consideration in this study. But we simulate the HVS and then focus on local pixel distortions that can occur from necessary stereoscopic image processing.

The primary visual cortex receives visual information coming from the eyes. After reaching the visual cortex, the human mind processes that sensory inputs and uses it to realize the scene. Image gradients provide important visual information which are essential for understanding the scene. Therefore, we believe that such information is important for the human visual system to understand the scene and judge its quality. This theory is supported by numerous FR IQA schemes based on the concept of gradient similarity. In relation to our problem, we use gradient magnitude and orientation as quality-aware features to evaluate the quality of stereoscopic images.

3.2.6 Gradient magnitude and orientation

Three gradient maps are produced from the cyclopean image, and disparity map using horizontal and vertical direction derivatives, d_x and d_y respectively. Gaussian distribution function is used as a kernel in a 5 by 5 mask to compute the directional gradient components [$d_x(i, j)$, and $d_y(i, j)$]. The mask weights are samples from 2D Gaussian function which is defined as follows:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\dfrac{x^2 + y^2}{2\sigma^2}} \tag{3.6}$$

where σ controls the amount of smoothing. If σ increases, more samples must be obtained to represent the Gaussian function accurately. The derivatives have been computed using central difference. In our implementation, we used a limited smoothing mask as it tends to extract more edge information which makes the gradients more sensitive to distortions. Thus, σ is fixed to 0.5 ($\sigma = 0.5$).

We compute for the obtained cyclopean image and disparity map, a Gradient magnitude (GM), Relative gradient Orientation (RO), and Relative gradient Magnitude (RM) Where the gradient magnitude is defined as:

$$|\nabla I(i, j)|_{GM} = \sqrt{d_x(i, j)^2 + d_y(i, j)^2} \tag{3.7}$$

while the gradient orientation is given by:

$$\angle \nabla I(i, j) = \arctan \frac{d_y(i, j)}{d_x(i, j)} \tag{3.8}$$

the relative gradient orientation is defined as follows:

$$\angle \nabla I(i, j)_{RO} = \angle \nabla I(i, j) - \angle \nabla I(i, j)_{AV} \tag{3.9}$$

where the local average orientation is:

$$\angle \nabla I(i, j)_{AV} = \arctan \frac{d_y(i, j)_{AV}}{d_x(i, j)_{AV}} \tag{3.10}$$

while the average directional derivative over x and y is defined by:

$$I_\gamma(i, j)_{AV} = \frac{1}{MN} \sum_{m,n} I_\gamma(i - m, j - n) \tag{3.11}$$

where M and N describe the size of the patches, 3×3 square neighborhood has been chosen (M=N=3), γ refers either to the horizontal x or the vertical y direction. Finally the relative gradient magnitude is defined by:

$$\begin{aligned} &|\nabla I(i,j)|_{RM} \\ &= \sqrt{(d_x(i,j) - d_x(i,j)_{AV})^2 + (d_y(i,j) - d_y(i,j)_{AV})^2} \end{aligned} \tag{3.12}$$

Standard deviation of each gradient histogram GM, RO, RM is computed as a final features extraction S_{GM}, S_{RO} and S_{RM}, respectively. The standard deviation is known as the square root of variance and defined by:

$$S(h) = \sqrt{\frac{1}{N-1}\sum_{x=1}^{N}(h(x) - \hbar)^2} \tag{3.13}$$

where $\hbar$ is the sample mean of the histogram $h(x)$ (normalized to unit sum), and N is the number of observations in the sample.

Display resolution is important factor for judging the quality. The subjective evaluation of a given stereo image varies when this factor changes. Therefore for objective evaluation, multi-scale method is a convenient way to incorporate image details at different resolutions. Wang *et al* [121] proposed a multi-scale quality assessment metric that outperforms the single-scale SSlM [120] model. The authors compared different down-sampling parameters results, and noted that down-sampling with a factor of 0.5 gives the best performance. Consequently, the cyclopean image is down-sampled with a factor of 0.5 (divided by 2), considering the changes in stereo image resolution and visual conditions. For example, distance from the viewer to the screen can change the size of the formed cyclopean view in his brain. The features S_{GM}, S_{RO} and S_{RM} are computed for each scale, yielding 6 features element from the cyclopean image. The final feature vector (F) has nine elements as follows:

$$\begin{aligned} &F = [S_{GM1}, S_{RO1}, S_{RM1}, S_{GM2}, S_{RO2}, S_{RM2}, \\ &S_{GMd}, S_{ROd}, S_{RMd}] \end{aligned} \tag{3.14}$$

Figure 3.6 illustrates the computed maps GR, RM, and RO from the cyclopean image over five well-known distortions. It can be observed that the distortions have affected differently the computed gradient maps GR, RM, and RO.

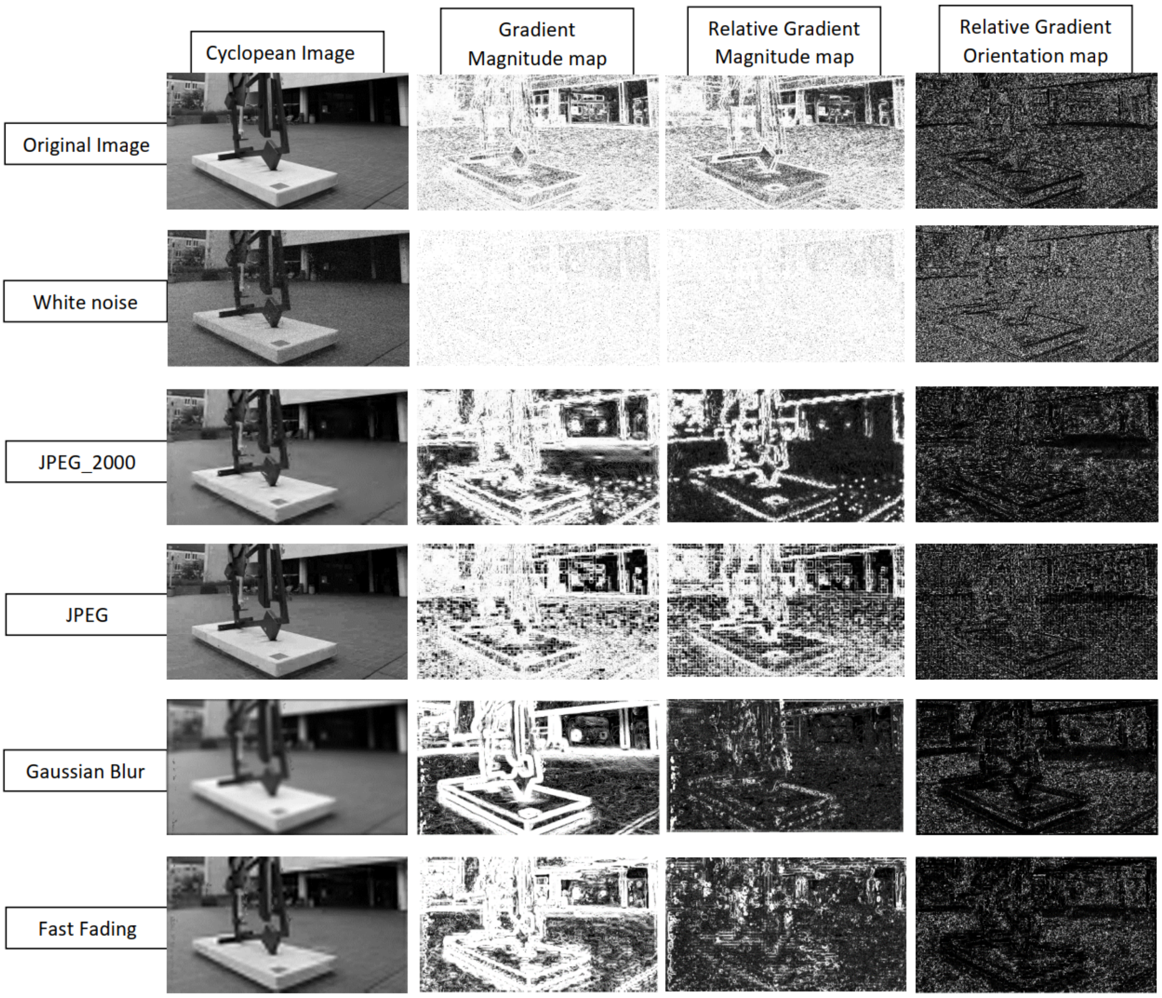


Figure 3.6: Examples of the constructed cyclopean image, GM, RM, and RO maps for different type of distortions.

Figure 3.7 displays the overview process to measure the quality of stereoscopic images, while Fig. 4.8 exhibits a 3D-plot of the extracted features over three databases. The represented features have been extracted from the cyclopean image in scale 1. The colored dots in the figure represent the extracted indicators in 3-dimensions. This 3D view shows that the features dots follow the same pattern on all databases which contain stereo images of different quality. Consequently, the extracted gradient indicators can be deployed for assessing the quality of stereoscopic images.

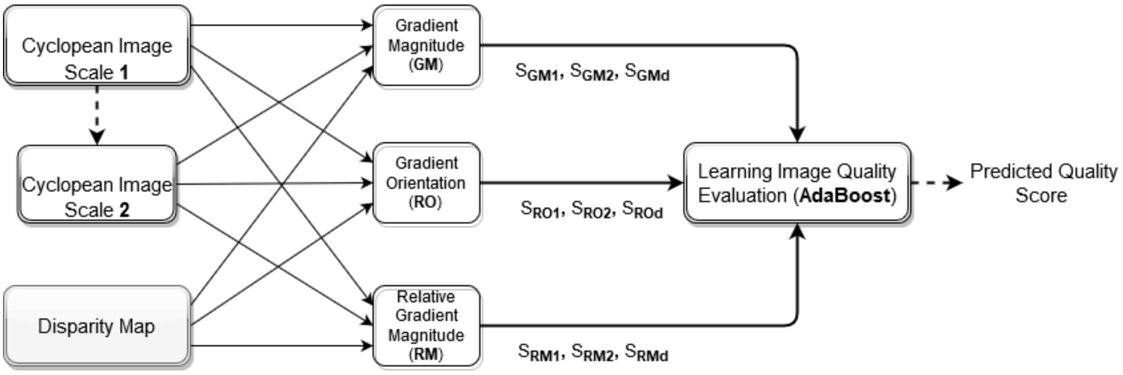

Cyclopean Image Scale **1**
Cyclopean Image Scale **2**
Disparity Map
Gradient Magnitude (GM)
Gradient Orientation (RO)
Relative Gradient Magnitude (RM)
S_GM1, S_GM2, S_GMd
S_RO1, S_RO2, S_ROd
S_RM1, S_RM2, S_RMd
Learning Image Quality Evaluation (**AdaBoost**)
Predicted Quality Score

Figure 3.7: Flowchart of the proposed measure.

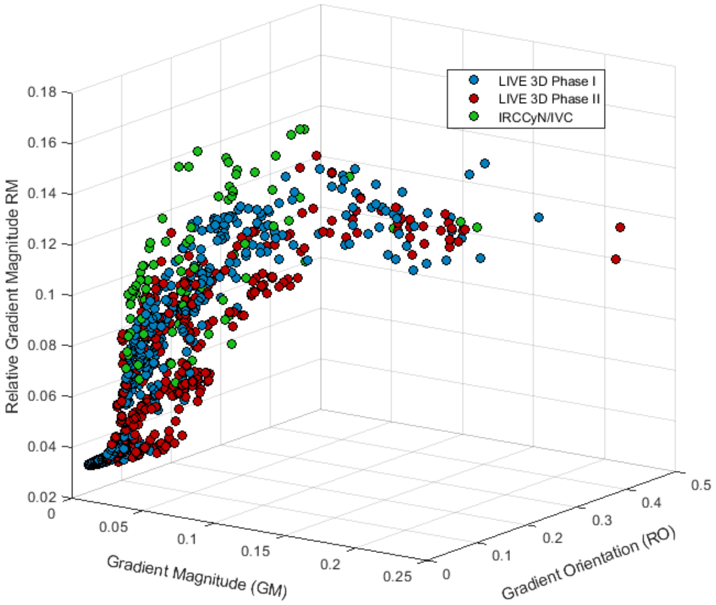


Figure 3.8: 3D-plot of the extracted features S_{GM1}, S_{RO1}, and S_{RM1} from the cyclopean image using LIVE 3D phase I, phase II and IVC 3D databases.

3.2.7 Learning for image quality evaluation: AdaBoost neural networks

Machine Learning (ML) plays an important role in the development of modern picture quality models. Although a limited number of IQA models have used advanced ML techniques such as AdaBoost [41]. The AdaBoost is an algorithm that consists in sequentially training a new simple model based on the errors of the previous model. A weight is assigned to each model. In the end, the whole set is combined to become an overall predictor. AdaBoost is one of the most useful ensemble method [10]. It can be used in conjunction with many other types of learning algorithm usually called Weak Learners (WL). The structure of the boosting ensemble generally outperforms a single feature learning model [93]. The boosting procedure tends to discover the examples and data points that are hard to predict and focuses on the next model predicting these examples better, by sequentially building a new simple model based on the errors of the previous model.

The use of Back-Propagation neural network is powerful for good prediction. Furthermore, to improve the performance of this neural network regression model, the AdaBoost idea has been implemented, and ANN with two hidden layers as WL has been deployed. However, the AdaBoost neural network can be less susceptible to the over-fitting problem than other learning algorithms. To solve this problem, 15% from training dataset has been dedicated validation for each neural network model.

The overall flow of the AdaBoost BP neural network model that computes the predicted output Q on a test set F is characterized as follows: First, set the quantity L of the Weak Learners (the BP artificial neural network models). Second, train the *ith* ANN on the sets X_j and Y_j, and estimate the predicted output of the testing set $Y_{i,j}^{pred}$. Afterward, a distribution D_i for the *ith* ANN is used for computing the evaluation error which is defined as (initial values of D_1 are set to 1):

$$D_{i+1,j} = D_{i,j} \times (1 + \delta . I(Y_j - Y_{i,j}^{pred})) \quad with \begin{cases} i = 1, .., L \\ \\ j = 1, .., M \end{cases} \tag{3.15}$$

The *ith* ANN evaluation error Err_i with the corresponding distribution D_i is defined as:

$$Err_i = \sum_{j=1}^{M} \mid D_{i,j} \times I(Y_j - Y_{i,j}^{pred}) \mid \tag{3.16}$$

where the function I is a binary function in which

$$I(x) = \begin{cases} 1 & \text{if } x > 0.2, \\ 0 & \textit{otherwise.} \end{cases} \tag{3.17}$$

Third, assign a weight w_i for the *ith* ANN using its error Err_i. Finally, the *ith* ANN predicts the quality P_i for the input F. For each ANN model, the adjusted weights and biases are randomly initialized. Hence, it produces varied L number of WL models with different prediction scores. Error threshold for the binary function I is set to 0.2. M is set to the vector size dedicated for testing. j indexes the *jth* element in a vector whose range is the integers between 1 and M. For instance, $D_{1,j}$ stands for the jth element in the vector D_1. δ is a constant multiplication factor, both of threshold and δ values are fixed to 0.2.

A convex function is used to convert the error of each ANN into its weight, in order to give the ANN models with a low error a high weight, and models with a high error a small weight. ω_i is the *ith* ANN weight, given as:

$$\omega_i = \frac{1}{e^{Err_i}} \tag{3.18}$$

The overall predicted measure is given by the weighted sum of the collection as:

$$Q = \sum_{i=1}^{L} \omega_i \times P_i \tag{3.19}$$

For the training dataset output, human scores are normalized in the form of DMOS to min-max normalization [0,1]. Hence, the range of the predicted measure values is from 0 to 1. The closer to 0 the better quality of the stereo image is. Algorithm 1 describes the developed AdaBoost regression algorithm.

Algorithm 1: Adaptive Boosting (AdaBoost) regression.

```
$L, F$   // $L$: the number of Weak learners (WL), $F$: stereo image features vector.
$Q$   // $Q$: the predicted quality.
Data: dataset for training and testing.
$n \longleftarrow 1$; $i \longleftarrow 1$;   // Initialization.
$Tr \leftarrow random(Data, 80\%)$
 $Te \leftarrow random(Data, 20\%)$   // Divide data randomly for training and testing.
$M \leftarrow size(Te)$   // Get the testing vectors size.
$D_1(1 : M) \longleftarrow 1$   // Initialize the first distributions.
for $n = 1 : L$ do
    $Ttr \leftarrow random(Tr, 85\%)$
     $Vtr \leftarrow random(Tr, 15\%)$   // Holdout 15% from train set for validation.
    $WL_n \leftarrow random(weights, biases)$
    $Train(WL_n, Ttr)$   // Train and validate the WL.
    $Terr(1 : M) \longleftarrow 0$
     $Err \longleftarrow 0$   // Reset the testing and evaluation error for each WL.
    $Terr \leftarrow Test(WL_n, Te)$   // Compute the testing error.
    for $i = 1 : M$ do
        if $(Terr(i) > 0.2)$ then
            $Err \leftarrow Err + D_n(i)$   // update the distribution $D_{n+1}(i)$ for next WL and compute the evaluation error of the $nth$ WL.
            $D_{n+1}(i) \leftarrow D_n(i) \times (1 + \delta)$
        else
            $D_{n+1}(i) \leftarrow D_n(i)$
    $w_n \leftarrow \frac{1}{e^{Err}}$
    $P_n \leftarrow WL_n(F))$   // Get the prediction of the $nth$ WL.
$Q \longleftarrow \sum_{n=1}^{L} w_n * P_n$   // Compute the final quality score.
```

The AdaBoost neural network has been used to predict the stereo image quality. Taking the handcrafted features from the disparity and cyclopean image as inputs. In the BP neural network, nine inputs cells have been deployed as the size of the final features vector (F) described in equation (3.14). Elements of the F vector are also mentioned in figure 6 as input elements for the ANN. Two hidden layers have been employed with nine neurons each. The applied transfer functions are tangent sigmoid and ReLU for the first and second hidden layers, respectively as shown in figure 3.9. A pure linear transfer function $f(x) = x$ has been used for a single node output layer. In hidden layers, a number of tests have been carried out using various activation functions. The tangent sigmoid and ReLU functions have been selected, for their best performance.

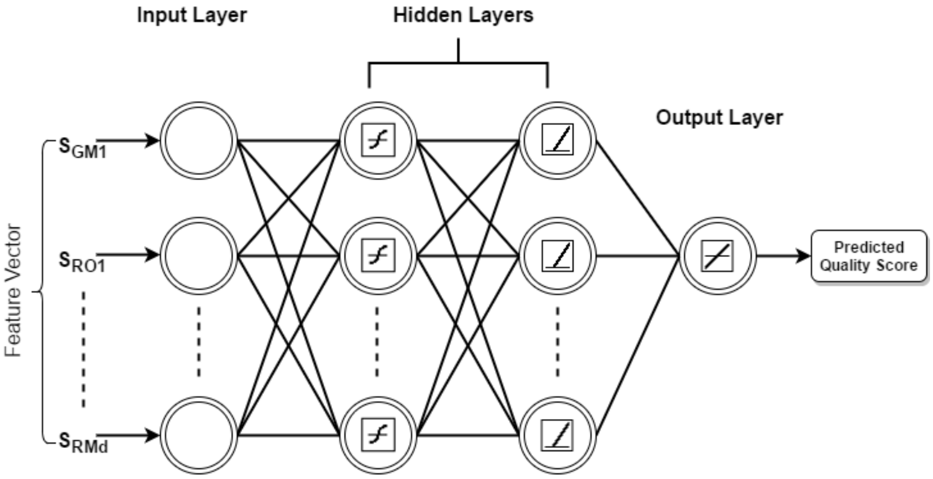


Figure 3.9: Structure of the used BP neural network.

3.2.8 Experiment protocols

The proposed approach has been tested on different databases. The obtained results have been compared to several FR and NR stereo IQA metrics, including six FR and eight NR stereo schemes. The standard performance assessment used in the Video Quality Experts Group (VQEG) has been considered. Objective scores are fitted to the subjective ones using logistic function in 1.4 previously discussed. Three widely-used performance indicators have been chosen to benchmark the proposed metric against the relevant state-of-the-art techniques: *LCC*, *SROCC* and *RMSE*. It is worth remembering that *LCC* and

RMSE assess the prediction accuracy while *SROCC* evaluates the prediction notability degree. Higher values for *LCC* and *SROCC* (close to 1) and lower values for *RMSE* (close to 0) indicate superior linear rank-order correlation and better precision with respect to human quality judgments, respectively. For a perfect match between the objective and subjective scores, $LCC = SROCC = 1$ and $RMSE = 0$.
Cross-validation training and testing provides a more accurate estimate of a model performance. However, several cross-validation techniques have been proposed, such as: LOOCV- Leave one out cross-validation and K-Fold cross-validation. The K-fold technique uses all data points to contribute to an understanding of how well the model performs the task of learning from some data and predicting some new data.

In order to ensure that the proposed approach is robust across content and it is governed by quality-aware indicators, the 5-fold cross validation over the three databases has been used. For every database, the dataset has been divided into 5 folds, where each fold contains a 80%-train set and 20%-test set randomly selected. The overlap between the test and the train set has been avoided to ensure that the reported results do not depend on features derived from known spatial information, which can artificially improve the performance. To demonstrate the generalization of the proposed metric against databases, a cross-database tests have been conducted. For further statistical performance analysis, a T-test scores have been computed over the correlation coefficients *LCC* and *SROCC*. In the different tests, the correlation of the feature vector with subjective human judgment has been studied. Complexity and time consuming of the proposed approach have been computed as well. Finally, influence of the formed cyclopean image and disparity map have been studied.

3.2.9 Feature vector correlation with human score

In this section, the feature vector F correlation with DMOS is evaluated. It is worth recalling that the regression model input is a vector of nine elements, and because of the restriction of human spatial awareness, it is hard to demonstrate the discriminative capacity of the characteristics in a graphical manner, such as a four-dimensional scatter plot. Three plots are used to describe the correlation of the adopted three indicators S_{GM}, S_{RO}, S_{RM} with the human opinion score.

Three-dimensional plots are used to visually depict the relationship between stereoscopic image quality and the three features. The extracted features are used as axes and each stereo image corresponds to a coordinate system scatter point. All the stereoscopic images from the LIVE 3D-I and LIVE 3D-II database are used for this demonstration. As shown in Fig. 3.10, the plots refer to features from cyclopean image scale 1, scale 2, and disparity map respectively from top to bottom. To differentiate the five types of distortion, we use distinct labels and map the DMOS rating of each stereo image to the preset color-map.

The ideal scenario is that the points are well separated with distinct kinds of distortion. It can be seen from Figure 9 that the scatter points of the five distortions are generally distinguished. The used stereo images are distorted increasingly from low to high factor. This can also be observed in the plots, where the adopted features vary smoothly in space with quality correspondence. Although there is some correlation between the extracted features, in particular S_{GM} and S_{RM}, where the coefficient correlation in terms of LCC is equal to 0.751. The deployed features provide good performance, this topic is discussed furthermore in section 3.2.13.

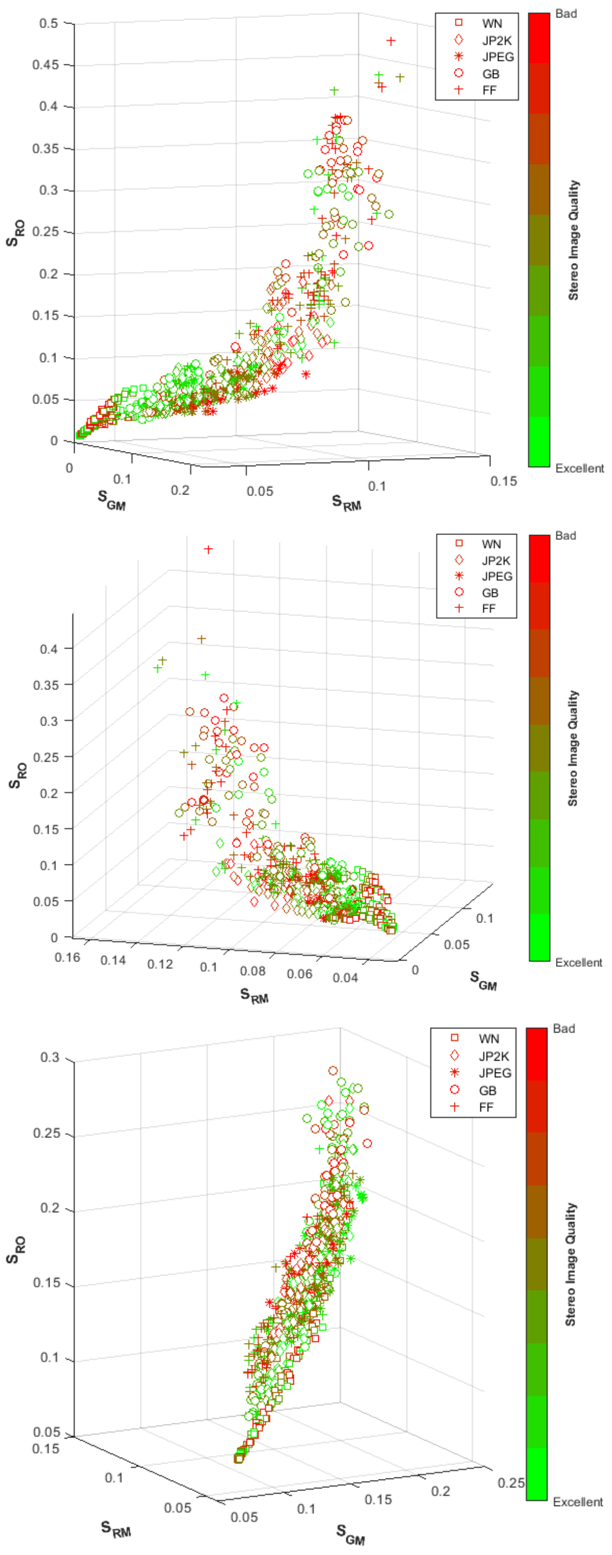


Figure 3.10: Illustration of the discriminatory power of the extracted features. Respectively from top to bottom: elements in the axis are from cyclopean image scale 1, scale 2, and disparity. (zoom in to get the markers more discriminative).

3.2.10 Comparison with other stereo IQA methods

The overall performance of the proposed scheme has shown good efficiency and consistency. The obtained results have been compared with several full-reference and no-reference stereo IQA metrics, including six FR, and eight NR metrics.

For the comparison purpose, two models have been created. The first model called *3D-nnet*, is a normal neural network regression model. It is equivalent to $L = 1$ in the AdaBoost algorithm (Algorithm 1). The second model named *3D-AdaBoost*, is a neural network combined with the AdaBoost technique as previously demonstrated, where 20 neural network models have been employed ($L = 20$). We find that the performance of the proposed measure is improved by using additional neural network models (Weak Learners) with saturation at a certain number and decreasing in the other case. Note that both models have the same network architecture. Also during the training, 15% is taken out from training set for validation. A Box plot in term of *SROCC* of the two models results is displayed in Fig. 3.11. Comparing the proposed models indicates that the performance can be improved by the adopted Adaptive Boosting technique.

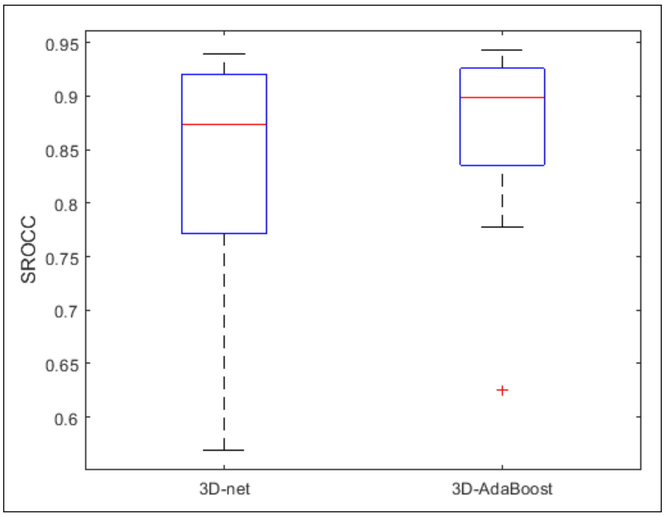


Figure 3.11: Comparison Box plots of *SROCC* of the proposed models. The *SROCC* results are split into four groups (quartiles). Each group has 25% of the results. The red line in the rectangle refers to the median value. Upper and lower ends of the rectangle limit the first and third quartiles, respectively. The length of the dashed line means the range of the mild outliers, and the symbol ”+” refers to the extreme outlier.

Tables 3.2, 3.3 and 3.4 show the results against DMOS of all stereo IQA algorithms on

LIVE 3D phase-I and phase-II. Furthermore, plots in the Figures 3.12 and 3.13 have been added to visualize the score responses on distortions separately.

Table 3.2: SROCC against DMOS on the LIVE 3D phase I II datasets.

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	0.930	0.910	0.603	0.931	0.699	0.899	0.923	0.751	0.867	0.455	0.773	0.728
You [130]		0.940	0.860	0.439	0.882	0.588	0.878	0.909	0.894	0.795	0.813	0.891	0.786
Gorley [45]		0.741	0.015	0.569	0.750	0.366	0.142	0.875	0.110	0.027	0.770	0.601	0.146
Chen [21]		0.948	0.888	0.530	0.925	0.707	0.916	0.940	0.814	0.843	0.908	0.884	0.889
Hewage [50]		0.940	0.856	0.500	0.690	0.545	0.814	0.880	0.598	0.736	0.028	0.684	0.501
Bensalma [13]		0.905	0.817	0.328	0.915	0.915	0.874	0.938	0.803	0.846	0.846	0.846	0.751
DIIVINE [83]	NR	-	-	-	-	-	0.882	-	-	-	-	-	0.346
Akhter [5]		0.914	0.866	0.675	0.555	0.640	0.383	0.714	0.724	0.649	0.682	0.559	0.543
Chen [22]		0.919	0.863	0.617	0.878	0.652	0.891	0.950	**0.867**	**0.867**	0.900	0.933	0.880
Lv [71]		-	-	-	-	-	0.897	-	-	-	-	-	0.862
Appina [7]		0.910	**0.917**	**0.782**	0.865	0.666	0.911	0.932	0.864	0.839	0.846	0.860	0.888
Zhou [135]		0.921	0.856	0.562	**0.897**	0.771	0.901	0.936	0.647	0.737	0.911	0.798	0.819
Fang [35]		0.883	0.880	0.523	0.523	0.650	0.877	**0.955**	0.714	0.709	0.807	0.872	0.838
3D-nnet		0.938	0.874	0.569	0.866	0.685	0.916	0.939	0.812	0.745	0.900	**0.934**	0.891
Proposed 3D-AdaBoost		**0.941**	0.899	0.625	0.887	**0.777**	**0.930**	0.943	0.842	0.837	**0.913**	0.925	**0.913**

Table 3.3: LCC against DMOS on the LIVE 3D phase I & II datasets.

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	0.925	0.939	0.640	0.948	0.747	0.902	0.926	0.784	0.853	0.535	0.807	0.784
You [130]		0.941	0.877	0.487	0.919	0.730	0.881	0.912	0.905	0.830	0.784	0.915	0.800
Gorley [45]		0.796	0.485	0.312	0.852	0.364	0.451	0.874	0.372	0.322	0.934	0.706	0.515
Chen [21]		0.942	0.912	0.603	0.942	0.776	0.917	0.957	0.834	0.862	0.963	0.901	0.907
Hewage [50]		0.895	0.904	0.530	0.798	0.669	0.830	0.891	0.664	0.734	0.450	0.746	0.558
Bensalma [13]		0.914	0.838	0.838	0.838	0.733	0.887	0.943	0.666	0.857	0.907	0.909	0.769
DIIVINE [83]	NR	-	-	-	-	-	0.893	-	-	-	-	-	0.442
Akhter [5]		0.904	0.905	0.729	0.617	0.503	0.626	0.772	0.776	0.786	0.795	0.674	0.568
Chen [22]		0.917	0.907	0.695	0.917	0.735	0.895	0.947	**0.899**	**0.901**	0.941	**0.932**	0.895
Lv [71]		-	-	-	-	-	0.901	-	-	-	-	-	0.870
Appina [7]		0.919	**0.938**	**0.806**	0.881	0.758	0.917	0.920	0.867	0.829	0.878	0.836	0.845
Zhou [135]		-	-	-	-	-	0.929	-	-	-	-	-	0.856
Fang [35]		0.900	0.911	0.547	0.903	0.718	0.880	**0.961**	0.740	0.764	0.968	0.867	0.860
3D-nnet		0.941	0.919	0.625	0.908	0.777	0.923	0.948	0.821	0.758	0.960	0.921	0.900
Proposed 3D-AdaBoost		**0.941**	0.926	0.668	**0.935**	**0.845**	**0.939**	0.953	0.835	0.859	**0.978**	0.925	**0.922**

Table 3.4: RMSE against DMOS on the LIVE 3D phase I II datasets.

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	6.307	4.426	5.022	4.571	8.257	7.061	4.028	6.096	3.787	11.763	6.894	7.490
You [130]		5.621	6.206	5.709	5.679	8.492	7.746	4.396	4.186	4.086	8.649	4.649	6.772
Gorley [45]		10.197	11.323	6.211	7.562	11.569	14.635	5.202	9.113	6.940	4.988	8.155	9.675
Chen [21]		5.581	5.320	5.216	4.822	7.837	6.533	3.368	5.562	3.865	3.747	4.966	4.987
Hewage [50]		7.405	5.530	5.543	8.748	9.226	9.139	10.713	7.343	4.976	12.436	7.667	9.364
Bensalma [13]		-	-	-	-	-	7.558	-	-	-	-	-	7.203
DIIVINE [83]	NR	-	-	-	-	-	7.301	-	-	-	-	-	10.012
Akhter [5]		7.092	5.483	**4.273**	11.387	9.332	14.827	7.416	6.189	4.535	8.450	8.505	9.294
Chen [22]		6.433	5.402	4.523	5.898	8.322	7.247	3.513	**4.298**	**3.342**	4.725	**4.180**	5.102
Appina [7]		6.664	4.943	4.391	6.938	9.317	6.598	4.325	5.087	4.756	6.662	6.519	7.279
Zhou [135]		-	-	-	-	-	6.010	-	-	-	-	-	6.041
Fang [35]		-	-	-	-	-	7.191	-	-	-	-	-	5.767
3D-nnet		5.622	5.083	5.104	6.059	7.819	6.277	3.394	5.598	4.780	3.889	4.481	4.905
Proposed 3D-AdaBoost		**5.593**	**4.867**	4.862	**5.104**	**6.633**	**5.605**	**3.226**	5.396	3.752	**2.859**	4.352	**4.352**

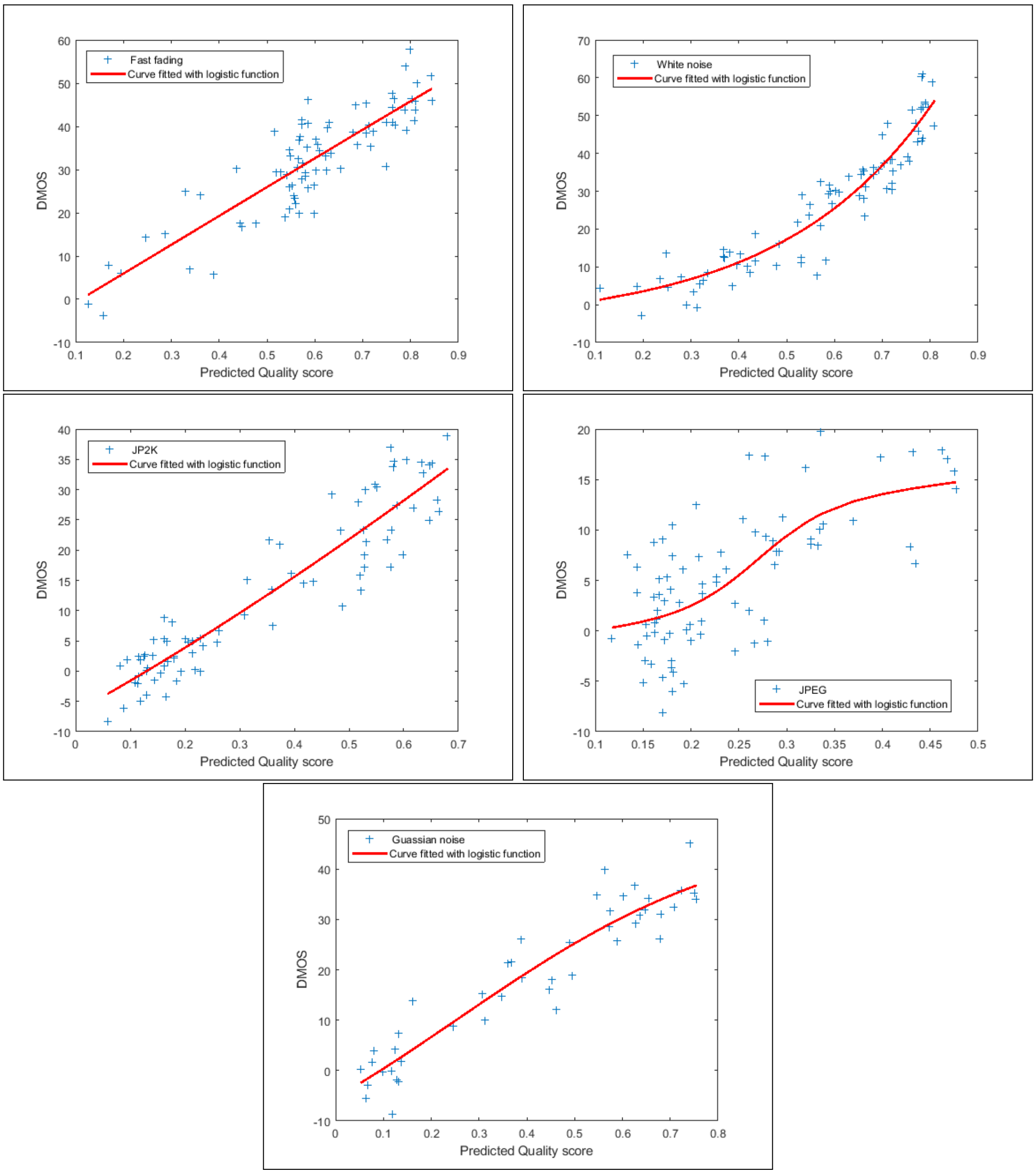


Figure 3.12: Scatter plot of the five distortions scores from LIVE 3D phase I IQA database using *3D-AdaBoost* method.

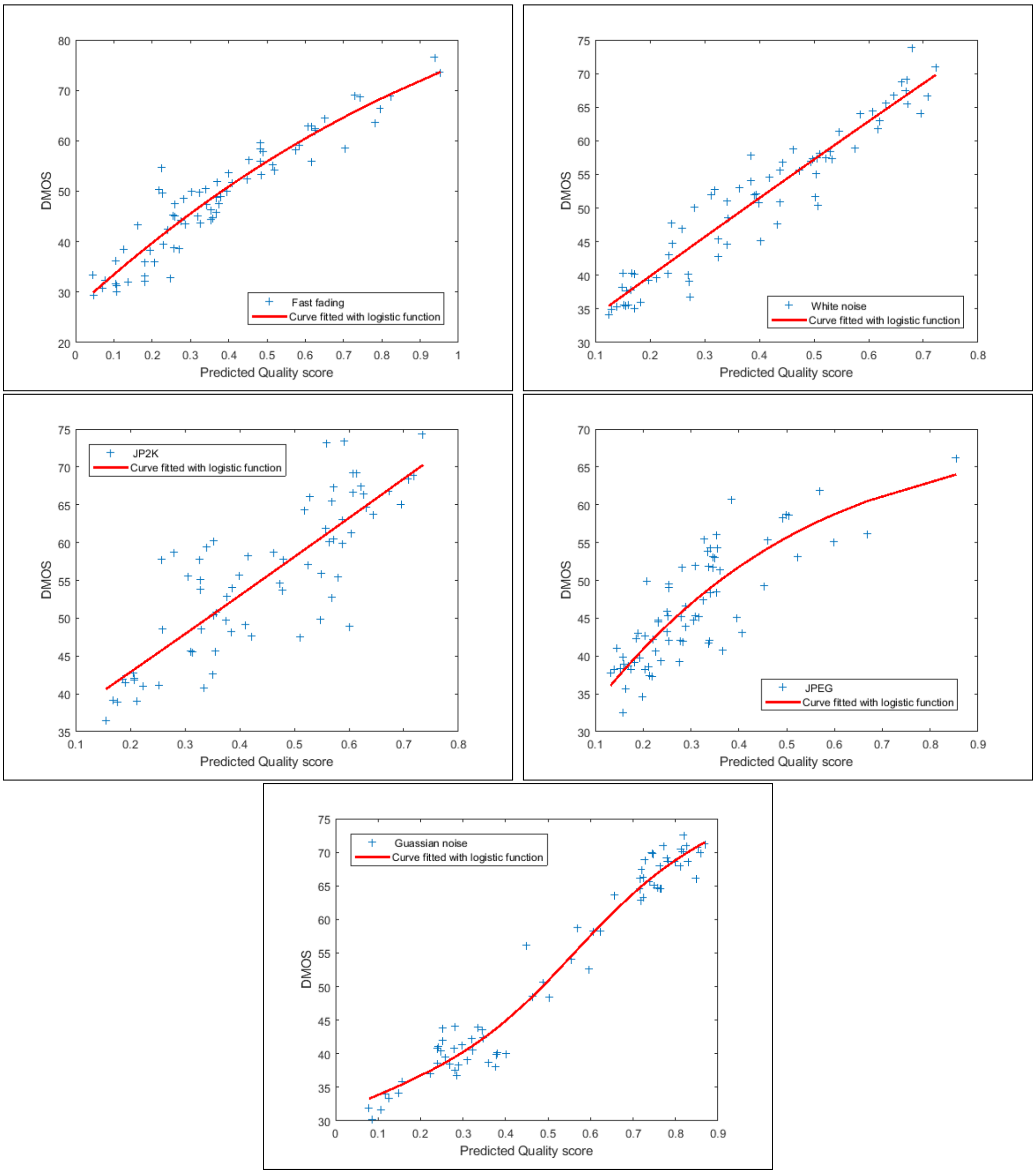


Figure 3.13: Scatter plot of the five distortions scores from LIVE 3D phase II IQA database using *3D-AdaBoost* method.

Moreover, the performance on symmetrically and asymmetrically distorted stimuli are shown separately in Table 3.5, while Table 3.6 provides detailed results over the five distortions. The 3D-nnet model has exhibited a good performance. Among all the comparison metrics, the model has obtained the best SROCC score on the two LIVE 3D databases, (SROCC=0.916 on LIVE 3D-I and SROCC=0891 on LIVE 3D-II). Finally, Table 3.7 shows the results on IVC 3D database. In these tables, the top NR methods results are highlighted in bold.

Table 3.5: SROCC result on Symmetric and Asymmetric distortion from LIVE 3D phase II dataset.

Method	*Type*	*Symmetric*	*Asymmetric*
Benoit [12]		0.860	0.671
You [130]		0.914	0.701
Gorley [45]	*FR*	0.383	0.056
Chen [21]		0.923	0.842
Hewage [50]		0.656	0.496
Bensalma [13]		0.841	0.721
DIIVINE [83]		–	–
Akhter [5]		0.420	0.517
Chen [22]		**0.918**	0.834
Lv [71]	*NR*	–	–
Appina [7]		0.857	0.872
Zhou [135]		–	–
Fang [35]		–	–
3D-nnet		0.861	0.902
Proposed 3D-AdaBoost		0.898	**0.917**

Table 3.6: Detailed results of SROCC, LCC, and RMSE on symmetric / asymmetric distortion from LIVE 3D-II.

Method	*Indicator*	*WN*	*JP2K*	*JPEG*	*Blur*	*FF*	*All*
Proposed 3D-AdaBoost	*SROCC*	0.923	0.829	0.933	0.848	0.889	0.898
Symmetric	*LCC*	0.938	0.922	0.946	0.913	0.903	0.903
	RMSE	3.701	3.709	3.819	3.425	4.876	4.609
Proposed 3D-AdaBoost	*SROCC*	0.897	0.926	0.897	0.921	0.945	0.917
Asymmetric	*LCC*	0.930	0.947	0.917	0.932	0.953	0.930
	RMSE	4.191	4.006	4.747	3.450	3.387	4.216

Table 3.7: SROCC, LCC, and RMSE against DMOS on the IVC 3D database.

Method	*Type*	*SROCC*	*LCC*	*RMSE*
Benoit [12]		–	–	–
You [130]		–	–	–
Gorley [45]	*FR*	–	–	–
Chen [21]		0.676	0.683	17.100
Hewage [50]		–	–	–
Bensalma [13]		–	–	–
DIIVINE [83]		0.422	0.486	18.259
Akhter [5]		–	–	–
Chen [22]		**0.851**	0.835	12.088
Lv [71]	*NR*	–	–	–
Appina [7]		–	–	–
Zhou [135]		–	–	–
Fang [35]		–	–	–
3D-nnet		0.780	0.779	13.830
Proposed 3D-AdaBoost		0.831	**0.845**	**11.776**

The proposed model *3D-AdaBoost* has given the best performance among all compared no-reference algorithms, while the full-reference method of Chen [21] yields better performance compared to other FR methods. Fig. 3.14 exhibits the prediction responses against human score DMOS on the three databases. Even though the proposed model is not designed for a specific distortion type. The comparison results on each individual distortion type indicate superiority of the proposed method over the three databases. More specifically, notice that the most of existing stereo IQA methods remain limited in capability and efficiency on asymmetric degradations. These metrics are more appropriate for symmetric distortion, but insufficient for asymmetric one. On the other hand, results show that the proposed framework delivers efficient performance over asymmetric/symmetric distortion. The method achieved LCC scores of 0.930 and 0.903 respectively on asymmetric/symmetric degradation. For better visualization, a scatter plots in Fig. 3.15 show the predicted quality on these two types of distortion separately (asymmetric/symmetric).

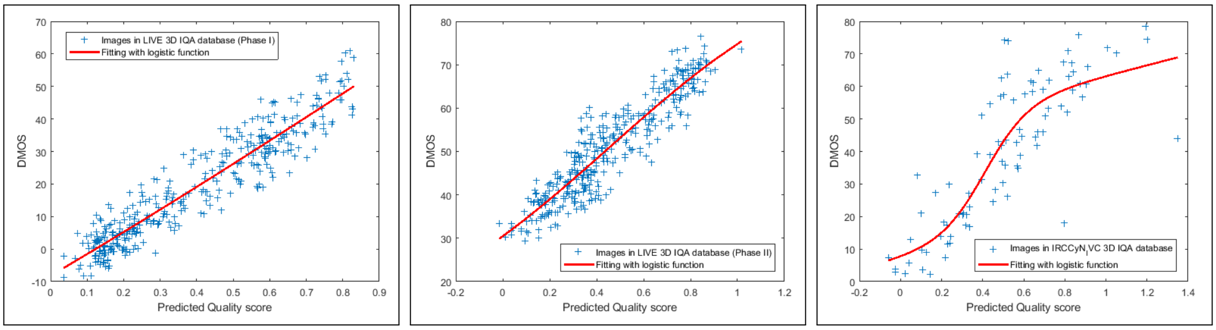


Figure 3.14: Scatter plots of subjective scores versus scores from the proposed scheme on the three stereopair IQA databases.

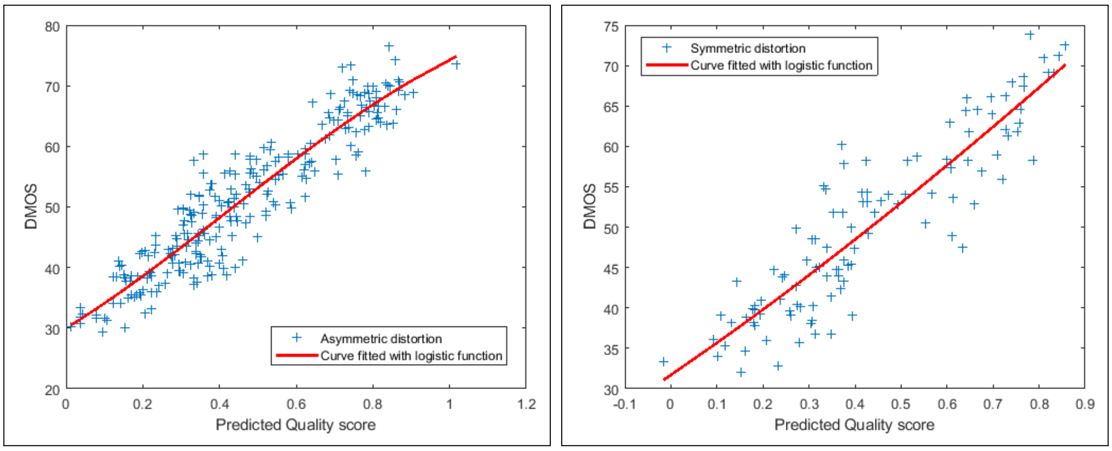


Figure 3.15: Scatter plot of asymmetric and symmetric distortions scores from LIVE 3D phase II IQA database using *3D-AdaBoost* method.

It should be noticed that the used neural network model presents better performance than the most commonly used (SVR) Support Vector Regression. The same evaluation process is followed for SVR with 5-fold cross validation. A radial basis function (RBF) kernel has been selected. The other parameters such as the number of support vectors and iterations are adjusted automatically during training for the best fit. Table 3.8 shows the superiority of the implemented neural network architecture over SVR. The mean scores of each learning method over three databases (LIVE 3D-I, LIVE 3D-II, and IVC 3D) have been calculated.

Table 3.8: Mean of SROCC, LCC, and RMSE results from the three databases using different regressors.

Method	*SROCC*	*LCC*	*RMSE*
SVR	0.8223	0.8406	9.0530
3D-nnet	0.8623	0.8673	8.3373
Proposed 3D-AdaBoost	**0.8913**	**0.9020**	**7.2443**

3.2.11 Performance using T-test

T-test is one of several types of statistical tests [91]. It questions whether the difference between the groups represents a true difference in the study or if it is likely a meaningless statistical difference, where 1 indicates that the groups are statistically different and 0 indicates that the groups are statistically similar. In order to investigate the statistical performance of the proposed metric, it is compared with the state-of-the-art methods. We conducted a left-tail T-test with confidence at 90% applied over 100 trials for PLCC and SROCC. The results provided in Table 5.14 show the superiority of the proposed method over the existing ones.

Table 3.9: T-test results with confidence of 90% of the proposed metric against the others using PLCC, SROCC from LIVE I and II

Method		*Akhter*	*Chen*	*Appina*	*Zhou*	*Fang*	*3D-nnet*
LIVE I	*LCC*	1	1	1	1	1	0
	SROCC	1	1	0	1	1	1
LIVE II	*LCC*	1	1	1	1	1	1
	SROCC	1	1	1	1	1	1

3.2.12 Cross-database performance

The above tests are useful for assessing robustness and generalization of the proposed metric, since all the results are obtained by training and testing using 5-fold cross validation. We extend cross-database experiments to demonstrate the performance capability of the proposed metric. The LIVE 3D phase I and phase II databases have been selected for these experiments because of the similarity in the number of stereo images. The model is trained on one database and tested on another one.

The Weak Learners (WL) in the *3D-AdaBoost* algorithm 1 are trained, validated, and tested on the LIVE 3D phase I database to obtain a model which will be tested on the LIVE 3D phase II database. Images in the LIVE 3D phase I database have been used for training, validating and testing, and images from the LIVE 3D phase II database are used as a final test set. The obtained results on LIVE 3D-1 are shown in Table 3.10 using LIVE 3D-II for training and LIVE 3D-I for test and vice versa. While Tables 3.11 presents detailed results over the five distortions. The *SROCC* has been used as a performance index. The best results are highlighted in bold.

Table 3.10: cross database SROCC, LCC, and RMSE results, Trained/Tested.

		L-II/L-I			L-I/L-II		
Method	Type	SROCC	LCC	RMSE	SROCC	LCC	RMSE
DIIVINE [83]	NR	0.882	0.893	7.301	0.346	0.442	10.012
Akhter [5]		0.383	0.626	14.827	0.543	0.568	9.294
Chen [22]		0.891	0.626	7.247	0.543	**0.895**	**5.102**
Lv [71]		0.897	0.901	-	0.862	0.870	-
Appina [7]		**0.911**	0.917	6.598	**0.888**	0.845	7.279
Zhou [135]		0.901	**0.929**	**6.010**	0.819	0.856	6.041
Fang [35]		0.877	0.880	7.191	0.838	0.860	5.767
3D-nnet		0.880	0.888	7.514	0.798	0.813	6.561
Proposed 3D-AdaBoost		0.887	0.897	7.224	0.823	0.832	6.253

Table 3.11: cross database SROCC results on the five distortions, Trained/Tested.

	L-II/L-I						L-I/L-II					
Method	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
DIIVINE [83]	-	-	-	-	-	0.882	-	-	-	-	-	0.346
Akhter [5]	0.914	0.866	0.675	0.555	0.640	0.383	0.714	0.724	0.649	0.682	0.559	0.543
Chen [22]	0.919	0.863	0.617	0.878	0.652	0.891	0.950	**0.867**	**0.867**	0.900	0.933	0.880
Lv [71]	-	-	-	-	-	0.897	-	-	-	-	-	0.862
Appina [7]	0.910	**0.917**	**0.782**	0.865	0.666	**0.911**	0.932	0.864	0.839	0.846	0.860	**0.888**
Zhou [135]	0.921	0.856	0.562	**0.897**	**0.771**	0.901	0.936	0.647	0.737	0.911	0.798	0.819
Fang [35]	0.883	0.880	0.523	0.523	0.650	0.877	**0.955**	0.714	0.709	0.807	0.872	0.838
3D-nnet	0.955	0.873	0.588	0.808	0.527	0.880	0.882	0.803	0.772	**0.925**	**0.936**	0.798
Proposed 3D-AdaBoost	**0.956**	0.889	0.556	0.875	0.530	0.892	0.932	0.826	0.737	0.881	0.924	0.824

It can be noticed that *3D-AdaBoost* trained on the LIVE 3D phase I database achieved lower performance compared to the model trained on phase II. This is due to the lack of asymmetric distortion in the LIVE 3D phase I database. However, it is interesting to observe that the *3D-AdaBoost* method produces good results on the LIVE 3D phase I database. Compared to other methods, although their results are not performed using cross-dataset test, the proposed metric ensures competitive performance on any type of distortion commonly encountered. Scatter plots in Fig. 3.16 show the *3D-AdaBoost* metric responses of cross-dataset test.

The overall experimental results have shown that the proposed method has good consistency among five distortion types with human subjective evaluation. The cross-database test showed the proposed metric reliability for measuring the quality of the stereoscopic image. Among the five distortions, JPEG distortion has the lowest accuracy. We believe this is due to the complexity of the compression distortion. Thus, it should be addressed separately for the stereo image quality assessment.

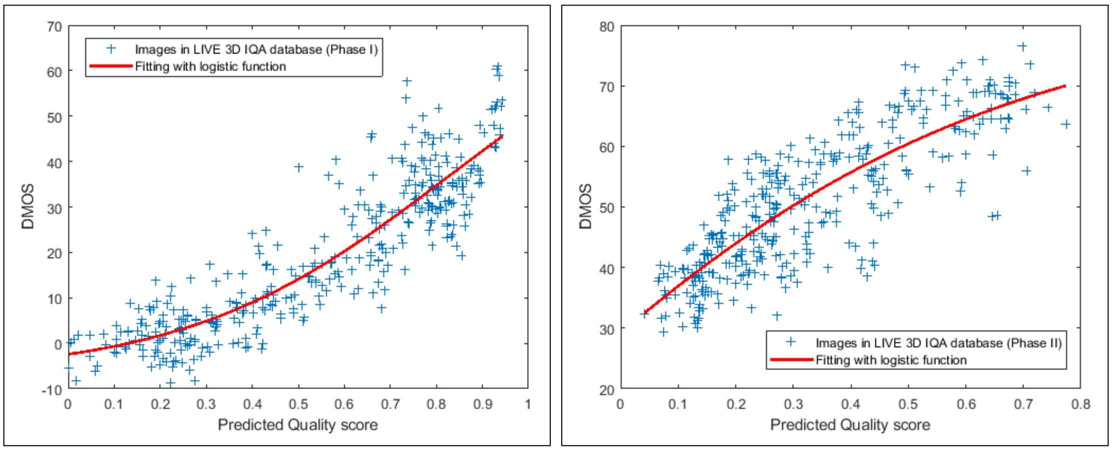


Figure 3.16: Scatter plot of cross-dataset scores on LIVE 3D phase I II databases using *3D-AdaBoost* method.

3.2.13 Influence of cyclopean view and disparity map

In order to demonstrate the efficiency of the proposed approach for measuring the stereo image quality, numerous tests have been conducted that cover the possibilities of feature extraction part. Also a simple feature extraction has been used for comparison. Pixel sum and pixel average have been used. The proposed learning part remains the same as described, using the 5-fold cross validation. The *3D-AdaBoost* model receives different input at each combination, and the mean performance of each combination is calculated over the three databases. The results of the tests are shown in Table 3.12. The pixel sum P_S is defined as follows:

$$P_S = \sum_{i=1}^{m} \sum_{j=1}^{n} I(i,j) \tag{3.20}$$

where I is the left or right image. The pixel average P_A is defined by:

$$P_A = \frac{1}{m.n} \sum_{i=1}^{m} \sum_{j=1}^{n} I(i,j) \tag{3.21}$$

Table 3.12: Mean of SROCC, LCC, and RMSE results from the three databases using various features and combinations.

The used material	*Features*	*N*	*SROCC*	*LCC*	*RMSE*
Stereopair Image	P_S	2	0.107	0.196	16.182
Stereopair Image	P_A	2	0.255	0.208	16.097
Stereopair Image	P_S, P_A	4	0.244	0.245	15.995
Stereopair Image	S_{GM}, S_{RO}, S_{RM}	3	0.740	0.769	10.386
Stereopair Image, disparity	S_{GM}, S_{RO}, S_{RM}	6	0.858	0.869	8.203
Stereopair Image, disparity	S_{GM}, S_{RO}, S_{RM}, P_S, P_A	12	0.716	0.741	9.803
Cyclopean view (scale 1)	S_{GM}, S_{RO}, S_{RM}	3	0.776	0.801	9.802
Cyclopean view (scale 1)	S_{GM}, S_{RO}, S_{RM}, P_S, P_A	5	0.676	0.705	10.552
Cyclopean view (scale 1, and 2)	S_{GM}, S_{RO}, S_{RM}	6	0.798	0.818	9.444
Cyclopean view (scale 1, and 2)	S_{GM}, S_{RO}, S_{RM}, P_S, P_A	10	0.688	0.715	10.298
Cyclopean view (scale 1, and 2), disparity	S_{GM}, S_{RO}, S_{RM}	9	**0.891**	**0.902**	**7.244**
Cyclopean view (scale 1, and 2), disparity	S_{GM}, S_{RO}, S_{RM}, P_S, P_A	15	0.628	0.635	12.445

From the results it can be observed that the pixel sum and average indicators give a bad performance, because these features do not correlate with the image quality. Meanwhile, the used features give good performance due to their relationship with distortion types and quality degradation as shown previously in Fig. 3.10. It is also noticeable that when using disparity map features, the performance improves which supports the study conducted in [9]. The authors used different measures to illustrated the relationship between the perceptual quality of stereo views and the quality of the disparity map. They concluded that the quality of the depth map is highly correlated with the overall 3D quality.

As discussed earlier, the 2D IQA metrics may not be applied to the stereo IQA problem, since either by averaging the score or the features obtained from left and right image will not consider asymmetrical distortions. The improved 2D IQA metric *DIIVINE* [83] for stereo images provides good performance on the LIVE 3D-I database and low performance on the LIVE 3D-II database. This is because the LIVE 3D-II database mainly contains asymmetric distorted stereo images. However, due to the fact that stereo images typically contain redundant information, a feature extraction from the left and right image may result in a redundant features. Therefore, the extracted features (S_{GM}, S_{RO}, and S_{RM}) from left and right images are averaged. Afterward, the *3D-AdaBoost* model has been used to map these features to predict the quality. It is also noticed that the use of 2-scale cyclopean image increases the accuracy of quality prediction. We assume that the space distance between cyclopean scale 1 features and cyclopean scale 2 features is learned while training, helping the model for better prediction. Additionally, the use of pixel sum P_S and pixel average P_A as features decreases the performance as shown in Table 3.12.

Even though the features are somewhat correlated, the model has given good results. Some tests have been carried out to support the use of all gradient extracted features (S_{GM}, S_{RO}, and S_{RM}). The performance deteriorates if one or two features among the three are neglected; as shown in Table 3.13. Therefore, it is important to utilize the three features for better quality assessment accuracy.

Table 3.13: Mean of SROCC, LCC, and RMSE results from the three databases using different gradient features and combinations.

The used material	*Features*	*N*	*SROCC*	*LCC*	*RMSE*
Cyclopean view (scale 1, and 2), disparity	S_{RO}	3	0.725	0.748	10.698
Cyclopean view (scale 1, and 2), disparity	S_{GM}	3	0.724	0.751	10.722
Cyclopean view (scale 1, and 2), disparity	S_{RM}	3	0.751	0.709	10.995
Cyclopean view (scale 1, and 2), disparity	S_{RO}, S_{RM}	6	0.809	0.817	9.401
Cyclopean view (scale 1, and 2), disparity	S_{GM}, S_{RM}	6	0.844	0.849	8.768
Cyclopean view (scale 1, and 2), disparity	S_{RO}, S_{GM}	6	0.838	0.847	8.823
Cyclopean view (scale 1, and 2), disparity	S_{GM}, S_{RO}, S_{RM}	9	**0.891**	**0.902**	**7.244**

Results given by Table 3.14 indicate that the cyclopean model using Gabor weights is better than the simple cyclopean model, in particular on the LIVE II. Compared to the stereo image model, the simple cyclopean model is also competitive, but the model may not be accurate on the asymmetric distortion situation as discussed earlier. The results of Table 3.15 also support the idea of using cyclopean view rather than using the stereo image for quality assessment problem. The superiority of the cyclopean image on symmetric and asymmetric degradations is also shown in Table 3.16. Notice that the performance of the stereopair image method in Table 3.15 drops significantly on LIVE 3D-II over all distortions. Consequently, for asymmetric distortions, extracting features directly from stereo images is not reliable. Also, the performance of cyclopean image method maintains consistency. In the Tables, N refers to the number of input features to the regression model *3D-AdaBoost*. Overall, we can conclude that the adopted cyclopean model and quality indicators (S_{GM}, S_{RO}, and S_{RM}) and the used combination are effective for assessing the quality of stereopair images.

Table 3.14: Cyclopean view versus Stereopair image method results over the three databases.

The used material	*Features*	N	*Indicator*	Live 3D-I	Live 3D-II	IVC 3D
Stereopair Image	S_{GM}, S_{RO}, S_{RM}	3	*SROCC*	0.905	0.725	0.590
			LCC	0.913	0.791	0.602
			RMSE	6.657	6.896	17.606
Cyclopean Image (scale 1)	S_{GM}, S_{RO}, S_{RM}	3	*SROCC*	0.908	0.797	0.622
			LCC	0.920	0.850	0.634
			RMSE	6.417	5.944	17.046
Cyclopean Image Simple (scale 1)	S_{GM}, S_{RO}, S_{RM}	3	*SROCC*	0.904	0.780	0.607
			LCC	0.914	0.828	0.622
			RMSE	6.644	6.323	17.263

Table 3.15: SROCC results of Cyclopean view versus Stereopair image method over LIVE 3D-I and LIVE 3D-II databases.

The used material	*Database*	*WN*	*JP2K*	*JPEG*	*Blur*	*FF*	*All*
Stereopair Image	LIVE 3D-I	0.943	0.867	0.597	0.816	0.615	0.905
	LIVE 3D-II	0.497	0.684	0.606	0.870	0.732	0.725
Cyclopean Image (scale 1)	LIVE 3D-I	0.943	0.869	0.589	0.867	0.680	0.908
	LIVE 3D-II	0.924	0.676	0.678	0.858	0.735	0.797
Cyclopean Image Simple (scale 1)	LIVE 3D-I	0.942	0.872	0.595	0.821	0.678	0.904
	LIVE 3D-II	0.911	0.706	0.674	0.844	0.727	0.780

Table 3.16: Cyclopean view versus Stereopair image method results on Symmetric and Asymmetric distortion from LIVE 3D-II dataset.

The used material	*Features*	*N*	*Indicator*	*Symmetric*	*Asymmetric*
Stereopair Image	S_{GM}, S_{RO}, S_{RM}	3	*SROCC*	0.672	0.745
			LCC	0.779	0.802
			RMSE	6.746	6.851
Cyclopean Image (scale 1)	S_{GM}, S_{RO}, S_{RM}	3	*SROCC*	0.733	0.822
			LCC	0.840	0.855
			RMSE	5.832	5.954
Cyclopean Image Simple (scale 1)	S_{GM}, S_{RO}, S_{RM}	3	*SROCC*	0.734	0.796
			LCC	0.814	0.834
			RMSE	6.249	6.337

3.2.14 Computational Complexity

Computational complexity of the proposed algorithm is discussed in this section. The most computationally expensive stage is the cyclopean image construction, since it involves weights computation of the left and right views by performing a multi-scale Gabor filter. The complexity of the proposed measure depends on the size of the testing vectors

(M) and the number of the Weak Learners (L). Therefore, the overall complexity of the proposed algorithm is *O(M . L)*. Furthermore, the computation time of the proposed model has been computed using a laptop computer with intel i5-2410M CPU, 2.30 GHz and 8 GB RAM, hence the run time in second is 72.5238 (including training time). There are no details on the complexity of the other NR methods. So, state-of-the-art metrics complexities have not been compared.

The stereoscopic image's pixel resolution may increase or decrease the run time, as well as the hardware computing power. The test has been conducted on the stereoscopic image shown in Fig. 3.1 of 640 x 360 pixels resolution. The more neural network models used, the higher the run time is. Clearly, the run time increases with the number of neurons. Note that the run time can be reduced via parallel computing (GPU cards) since the proposed method is based on neural networks.

3.2.15 Conclusion

A new blind stereoscopic IQA metric has been proposed. The model is based on human binocular perception and advanced machine-learning algorithm. Efficient perceptual features have been extracted from the gradient magnitude (GM) map, relative gradient orientation (RO) map and the relative gradient magnitude (RM) map. In the following, few points are concluded:

- Experimental results showed that the extracted features are sensitive to the five common distortions. Considering the variations of stereo image resolution and viewing conditions, a multi-scale gradient maps of the cyclopean image have been employed.
- AdaBoost neural network is used to map the stereo image features to quality score. The overall obtained results have indicated that the metric correlates well with subjective scores DMOS over symmetric and asymmetric distortions.
- The proposed metric performs better in terms of both accuracy and efficiency on the three publicly available stereoscopic IQA databases, LIVE 3D-I, LIVE 3D-II, and IRCCyN/IVC 3D than the state-of-the-art methods.

The use of the extracted features can also be useful for the development of no-reference stereoscopic video quality models. Furthermore, the proposed workflow allows to develop the idea of AdaBoost by incorporating other feature-learning algorithms. However, despite the good performance that AdaBoost technique offers, it significantly increases the runtime, but this latter can be compensated using parallel computing (e.i deploying GPU).

3.3 2nd approach: Automatic Distortion Type Recognition for Stereoscopic Images

3.3.1 Introduction

In the last decade, great efforts have been dedicated to the development of quality assessment and enhancement algorithms. But only few stereoscopic IQA metrics that use distortion classification have been proposed. For instance, authors in [59] have proposed ParaBoost (parallel-boosting) stereoscopic image quality assessment (PBSIQA) system. Their method firstly classifies the distortion type of the stereo image, and then multiple quality models are used to evaluate the stereo image quality. As discussed earlier, the SIQA metric in [38] uses strategy of distortion type identification. Where it determines whether the distortion is symmetric or asymmetric to account for the binocular fusion properties. However, despite the fact that distortion detection is useful for enhancing IQA metrics, it has received little consideration. For instance, it can be employed to select the image enhancement algorithm according to the distortion type as demonstrated in Fig. 3.17.

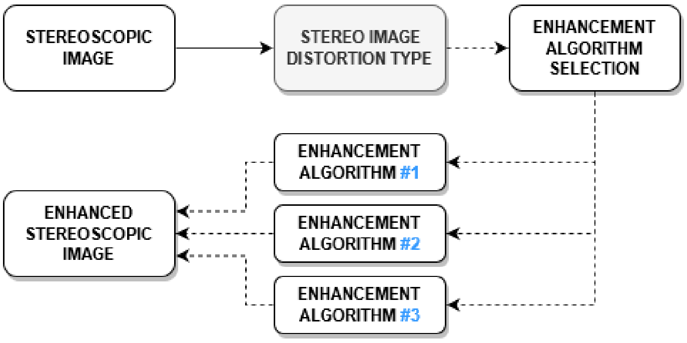


Figure 3.17: An application example of the proposed recognition system.

The aim of our approach is to design a no-reference classifier for stereoscopic image distortions, the model can then be used to develop important image processing algorithms, such as image quality assessment and restoration. The problem of distortion recognition in stereoscopic images is interesting and straight forward task. Since this classification model is the first of its kind that primarily addresses the problem, the recognition model has not been compared with related work.

3.3.2 Approach overview

In order to design a model that can assess or enhance the quality of stereoscopic images, the first step is to identify the distortions which could arise when dealing with stereoscopic content. In this work, the model involves three steps as idicated in Fig.3.18: first is construction of disparity map (as done in our previous approach). Second, gradient features are extracted from each view (left and right views) and disparity map. Third, predicting the quality based on feature learning, using SVM fitted model.
We handcrafted the same features as done in the proposed metric discussed earlier. Whereas three Gradient maps are computed over horizontal and vertical derivatives from each left, right image, and disparity. Afterwards, three indicators are extracted from these gradient maps: the gradient magnitude (GM), relative gradient orientation (RO), and the relative gradient magnitude (RM) as shown in Fig.3.18.

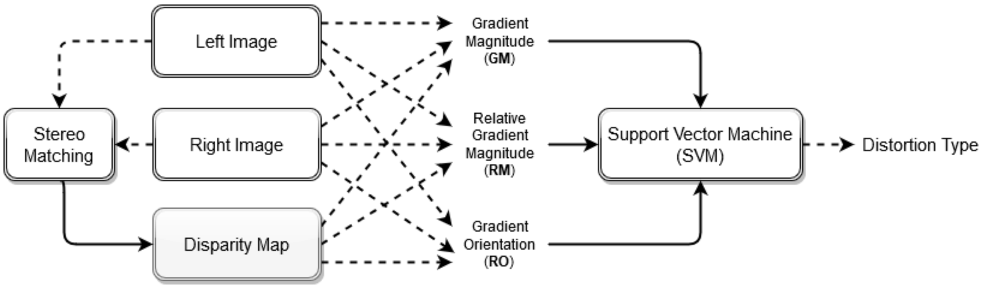


Figure 3.18: The proposed distortion type classification scheme.

3.3.3 Learning for distortion type recognition using SVM

The objective of the learning-based classifier is to learn how to map inputs X to out-put classes N. For our classification problem, a multi-class vector support (SVM) algorithm is used to predict stereo image distortion among other classifiers due to its good performance. It is applied on the extracted features dataset obtained from the previous steps. For the best fit, a Radial Basis Function (RBF) has been used as a kernel function in the SVM model. Fig. 3.19 shows the adopted simple workflow for quality prediction.

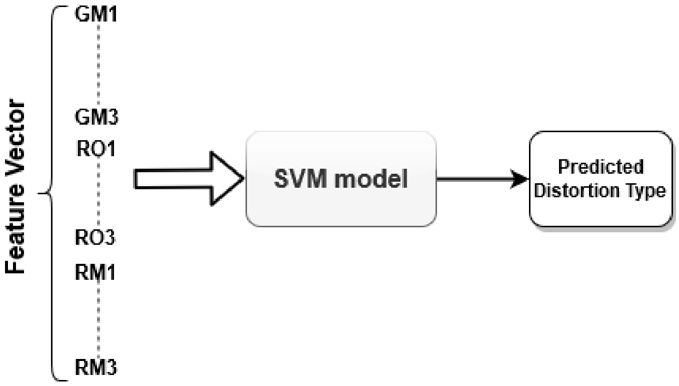


Figure 3.19: Prediction scheme using the extracted feature vector.

Where GM1, RO1, and RM1 are features retrieved from the left image. The retrieved features from the right image are GM2, RO2, and RM2. The retrieved features from the disparity map are GM3, RO3, and RM3.

3.3.4 Experimental results and analysis

To assess the efficiency of the suggested algorithm, the two databases were used. The well-known LIVE 3D phase I and phase II databases are used to show the performance of the proposed model. The two phases together combine the biggest and most comprehensive stereoscopic image quality database presently available.

In this work, 5-fold cross validation technique is used to evaluate the performance of the proposed method. The two stereoscopic image databases have been used to collect training and test set dataset, where the data set contains a total of 725 stereo images. Each fold is divided into 80% train set, and a 20% test set. The training and testing are conducted five times, this process guarantees that each data point ends up in the 20% test set exactly once. The model has achieved good performance among the five distortions. The best classification result has obtained on the WN degradation. That is because proposed classification scheme relies on gradient as features. Meanwhile, the JP2K degradation has worse classification performance due to the complexity of JP2K compression algorithm. Confusion matrices have been computed as shown in Fig. 3.20 to find out how the classifier performed in each class. Table 3.17 shows numbers associated with distortion types. Note that this study is the first work dealing with distortion types recognition in stereo images. Therefore, no comparisons with standards or related works are made.

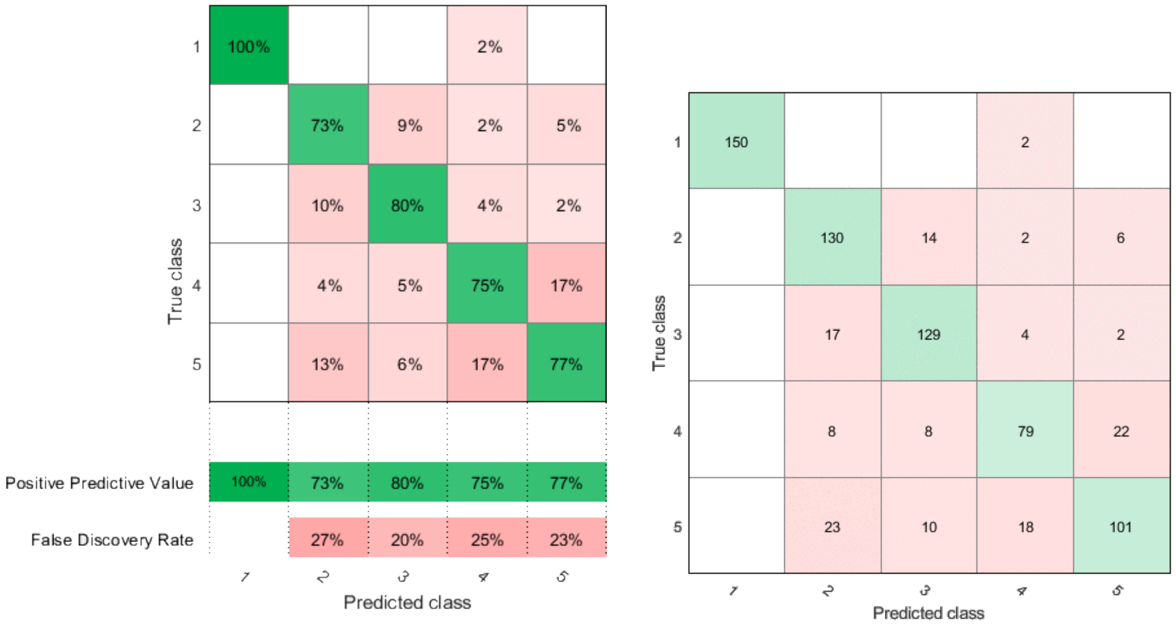
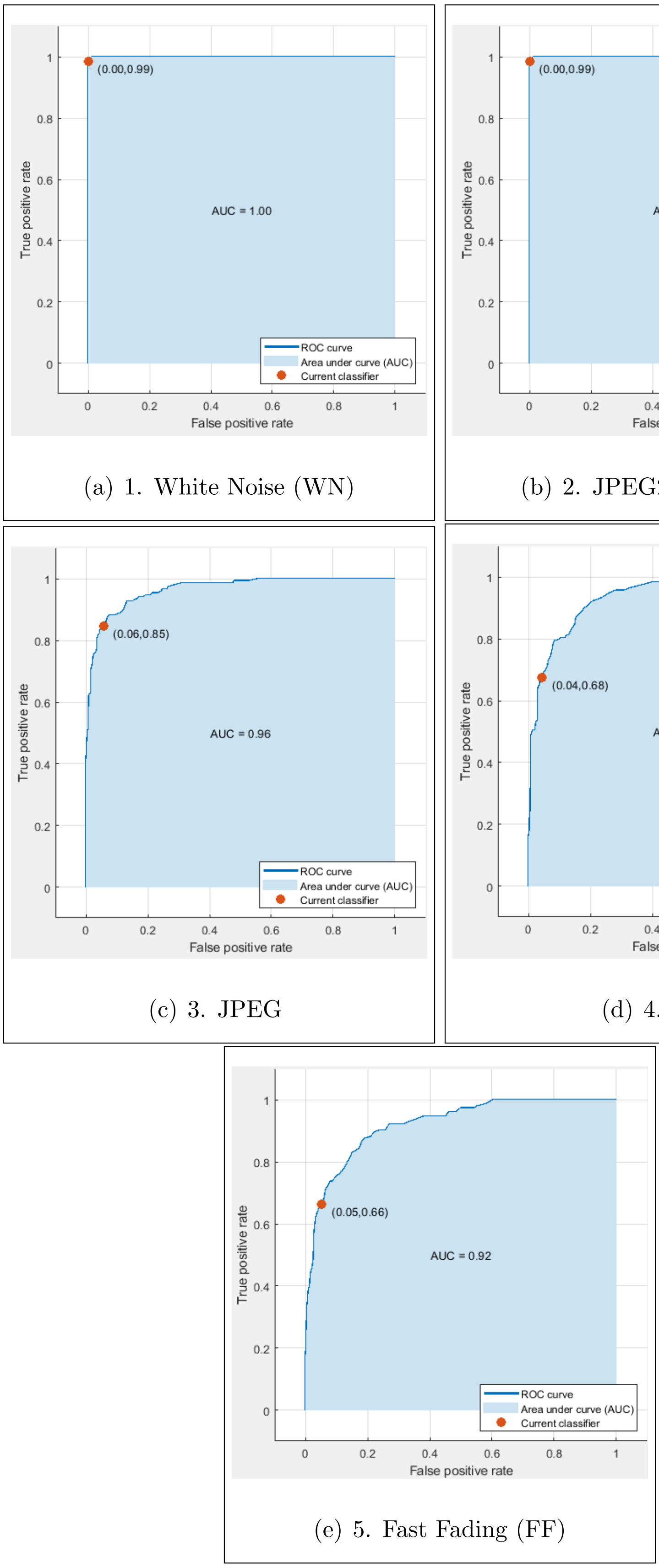


Figure 3.20: Confusion matrix plot using accuracy/number of observations of the proposed model.

Table 3.17: Distortions with their referring numbers.

Distortion type	*Referring Number*
White Noise (WN)	1
JPEG2000 (JP2K)	2
JPEG	3
Blur	4
Fast Fading (FF)	5

Furthermore, Receiver Operating Characteristic (ROC) curves in Fig. 3.21 show the performance at all classification thresholds. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold parameters. The area under the ROC curve, or AUC (Area Under Curve) varies from 0 to 1, AUC = 1 corresponds to a perfect classifier. Larger AUC values indicate a good classifier performance. The experimental results demonstrate that the proposed classifier has obtained significant classification consistency and accuracy over five types of distortion (WN, JP2K, FF, Blur).



(a) 1. White Noise (WN)

(b) 2. JPEG2000 (JP2K)

(c) 3. JPEG

(d) 4. Blur

(e) 5. Fast Fading (FF)

Figure 3.21: ROC curves of the proposed classifier over the five type of distortion.

3.3.5 Conclusion

Inspired by our 3D-Adaboost metric, a new NR distortion classification algorithm dedicated for stereoscopic images. The method enables each distortion to be addressed separately. Therefore, it can be used to improve algorithms for quality evaluation / restoration. The models are based on gradient information and machine learning. Magnitude and oriented gradient features are used to classify the distortion of stereoscopic image. SVM (learning based) model has been used to map the features to distortion type. The results obtained have shown that the system is reliable over the most common types of distortion. Furthermore, this scheme can be extended to determine whether a stereopair image is symmetrically or asymmetrically distorted and even measuring the degree of degradation.

Chapter 4

Contributions based on automatic quality feature extraction and HVS modeling

4.1 Introduction

Recently, the SIQA metrics have incorporated new machine learning techniques. Among these techniques, Deep Learning (DL) is the popular one being used for its ease of use and high performance, as mentioned in previous chapters. As a result, in this chapter, we explore DL algorithms for SIQA metrics. We present two DL-based SIQA approaches. One approach is built on an end-to-end CNN model [79], while the second approach is based on deep features bank, extracted automatically by the model [77]. In addition, we investigate the behavior and performance of the DL-based metrics.

4.2 1st model: NR-SIQA based on Deep learning and cyclopean view

4.2.1 Approach overview

In order to design a DL based SIQA model, we follow the previously used mathematical human binocular perception simulation. Therefore, the hypothesis of a cyclopean image is maintained while designing our metric. The model in this work involves three steps: the first step, Gabor filter responses, and disparity map have been used to construct the cyclopean image (see Fig. 3.4). Secondly, the cyclopean image has been divided into four patches to train four CNN models. Thirdly, the quality scores are predicted from the CNN models, and the scores average is computed.

4.2.2 Quality assessment: Deep Learning

In recent years, deep learning has been deployed to solve difficult problems such as image classification and speech recognition. The end-to-end network allows to extract automatically relative features from the raw data showing significant accuracy improvement in the IQA domain. The handcraft quality-aware features extraction is sometimes difficult, time-consuming, and requires expert knowledge, especially in quality evaluation. Therefore, using deep learning algorithm may solve these difficulties. A convolutional neural network is one of the most popular algorithms for deep learning. To build our SIQA model, four CNNs have been implanted in order to estimate the quality at each corner of the scene.

In previous work (presented in chapter 3), relative gradient features have been extracted from the cyclopean image, then an AdaBoost (Adaptive Boosting) neural network model has been created. The model that can predict the quality from the input features. However, in this work using deep learning will skip the step of manual features extraction. The cyclopean image is fed directly to the CNN models, which then predict the quality.

4.2.3 Network architecture

The architecture of the used CNN consists of 10 layers, four patches from the cyclopean image are all have a size of 180 × 320 pixels. Every CNN is trained and used to predict the quality score for each patch. After-all, the average score is computed from the four CNNs as shown in Fig. 4.1.

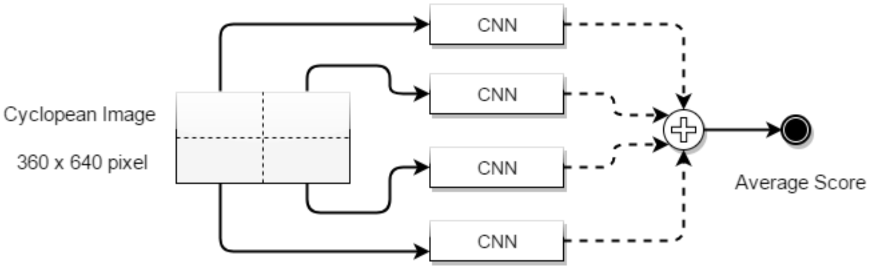


Figure 4.1: The flowchart of the proposed measure using four Convolutional Neural Networks. The cyclopean image divided into four patches, equally have a size of 180 × 320 pixels.

After each convolution layer, a ReLU layer is applied, followed by Max-pooling Layer. The used two convolutional layers have 10 filters each, with a size 3 × 3 and a stride of 1 pixel. The Max-pooling layer reduces the size of each feature map. It is achieved by applying a max filter. The max filter takes the maximum pixel value of a region. The first used Max-pooling layer has a size of 3 × 3 and a stride of 1 pixel. The second Max-pooling layer has a size of 8 × 8 and a stride of 8 pixels. A fully connected layer and regression layer with 1 node respectively come after the second Max-pooling. The architecture of the CNN is illustrated in Fig. 4.2.

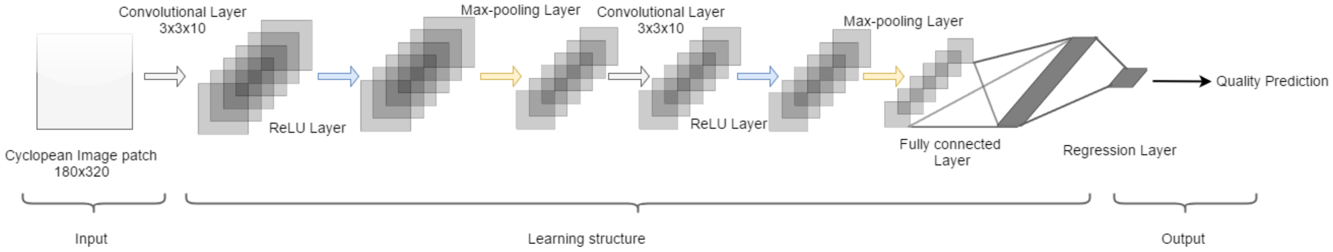


Figure 4.2: The proposed Convolutional Neural Network Architecture.

For the training dataset output, the human scores are normalized in the form of DMOS to min-max normalization [0,1]. Hence, the range of predicted score values is from 0 to

1. The closer to 0 the better quality of the stereo image is.

The four networks are trained for 100 epochs each and to prevent over-fitting, 15% from the dataset has been selected randomly as a validation set. In each epoch during the training, the *RMSE* is computed for both the validation and the whole dataset. The Back-Propagation algorithm updates the network parameters (filter weights) to minimize the *RMSE*. It is worth to remember that the RMSE is defined as follows:

$$RMSE = [\frac{1}{N}\sum_{n=1}^{N}(P_n - Q_n)^2]^{\frac{1}{2}} \tag{4.1}$$

where N is the number of image-patches input. P is the predicted score and Q is the normalized human objective score.

4.2.4 Experimental results and analysis

The popular LIVE 3D phase I and phase II databases are used to train and test the performance of the proposed metric. The performance has been evaluated via the three common indexes: The *RMSE*, *SROCC*, and the *LCC*. Higher values for *LCC* and *SROCC* (closer to 1) and lower values for *RMSE* (closer to 0) indicate superior linear rank-order correlation and better precision with respect to human quality judgments, respectively.

We examine the generalization capability and robustness of our metric by cross-dataset training and testing. The model is trained on such database and tested on another database. We have generated three 3D-CNN models: first model (3D-CNN$_{all}$) is trained and tested on both phase I and phase II. The second model (3D-CNN$_{live\ I}$) is trained on phase I and then tested on phase II. The third model (3D-CNN$_{live\ II}$) is a reverse of the 2nd model, it is trained on phase II and tested on phase I. In these cross-dataset tests, we train each one using 80% of database images, which are randomly selected. The remaining 20% of images are used as the validation set.

Tables 4.1, 4.2 and 4.3 exhibit the results against DMOS of all stereo IQA algorithms on LIVE 3D phase I and phase II. In the following, Table 4.4 contains the performance on symmetric and asymmetric distorted stimuli. The proposed model has given the better performance compared to most other metrics. In all tables, the top three methods result are highlighted in bold. A scatter plots of objective scores against subjective scores on LIVE 3D phase I and phase II are given in Fig. 4.3.

Table 4.1: SROCC against DMOS on the 3D LIVE Phase-I and Phase-II datasets.

	LIVE I						LIVE II					
Method	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	**0.923**	0.751	**0.867**	0.455	0.773	0.728	0.923	0.751	**0.867**	0.455	0.773	0.728
You [130]	0.909	**0.894**	0.795	0.813	**0.891**	0.786	0.909	**0.894**	0.795	0.813	**0.891**	0.786
Gorley [45]	0.875	0.110	0.027	0.770	0.601	0.146	0.875	0.110	0.027	0.770	0.601	0.146
Chen [21]	**0.940**	0.814	**0.843**	**0.908**	0.884	**0.889**	**0.940**	0.814	**0.843**	**0.908**	0.884	**0.889**
Hewage [50]	0.880	0.598	0.736	0.028	0.684	0.501	0.880	0.598	0.736	0.028	0.684	0.501
Bensalma [13]	0.905	0.817	0.328	**0.915**	**0.915**	0.874	**0.938**	0.803	0.846	0.846	0.846	0.751
Akhter [5]	0.714	0.724	0.649	0.682	0.559	0.543	0.714	0.724	0.649	0.682	0.559	0.543
Zhou [135]	0.921	**0.856**	0.562	0.897	0.771	**0.901**	0.936	0.647	0.737	**0.911**	0.798	0.819
Proposed *3D-CNN*$_{all}$	**0.976**	**0.927**	**0.852**	**0.951**	**0.973**	**0.964**	**0.970**	**0.955**	**0.871**	**0.909**	**0.927**	**0.944**
Proposed *3D-CNN*$_{live\ II}$ *and 3D-CNN*$_{live\ I}$	0.905	0.760	0.318	0.885	0.801	0.866	0.924	**0.920**	0.720	0.850	**0.889**	**0.878**

Table 4.2: LCC against DMOS on the 3D LIVE Phase-I and Phase-II datasets.

	LIVE I						LIVE II					
Method	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	**0.926**	0.784	**0.853**	0.535	0.807	0.784	**0.926**	0.784	**0.853**	0.535	0.807	0.784
You [130]	0.912	**0.905**	0.830	0.784	**0.915**	0.800	0.912	**0.905**	0.830	0.784	**0.915**	0.800
Gorley [45]	0.874	0.372	0.322	**0.934**	0.706	0.515	0.874	0.372	0.322	0.934	0.706	0.515
Chen [21]	**0.957**	0.834	**0.862**	**0.963**	**0.901**	**0.907**	**0.957**	0.834	**0.862**	**0.963**	**0.901**	**0.907**
Hewage [50]	0.891	0.664	0.734	0.450	0.746	0.558	0.891	0.664	0.734	0.450	0.746	0.558
Bensalma [13]	0.914	0.838	0.838	0.838	0.733	0.887	0.943	0.666	0.857	0.907	0.909	0.769
Zhou [135]	-	-	-	-	-	**0.929**	-	-	-	-	-	0.856
Akhter [5]	0.772	0.776	0.786	0.795	0.674	0.568	0.772	0.776	0.786	0.795	0.674	0.568
Proposed *3D-CNN*$_{all}$	**0.981**	**0.967**	**0.879**	**0.974**	**0.968**	**0.974**	**0.969**	**0.959**	**0.902**	**0.976**	**0.950**	**0.948**
Proposed *3D-CNN*$_{live\ II}$ *and 3D-CNN*$_{live\ I}$	0.911	**0.855**	0.433	0.910	0.818	0.885	0.919	**0.924**	0.720	**0.957**	0.903	**0.887**

Table 4.3: RMSE against DMOS on the 3D Live Phase-I and Phase-II datasets.

	LIVE I						LIVE II					
Method	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	**4.028**	6.096	**3.787**	11.763	6.894	7.490	**4.028**	6.096	**3.787**	11.763	6.894	7.490
You [130]	4.396	**4.186**	4.086	8.649	**4.649**	**6.772**	4.396	**4.186**	4.086	8.649	**4.649**	6.772
Gorley [45]	5.202	9.113	6.940	**4.988**	8.155	9.675	5.202	9.113	6.940	4.988	8.155	9.675
Chen [21]	**3.368**	**5.562**	**3.865**	**3.747**	**4.966**	**4.987**	**3.368**	5.562	**3.865**	**3.747**	4.966	**4.987**
Hewage [50]	10.713	7.343	4.976	12.436	7.667	9.364	10.713	7.343	4.976	12.436	7.667	9.364
Bensalma [13]	-	-	-	-	-	7.558	-	-	-	-	-	7.203
Akhter [5]	7.416	6.189	4.535	8.450	8.505	9.294	7.416	6.189	4.535	8.450	8.505	9.294
Zhou [135]	-	-	-	-	-	6.010	-	-	-	-	-	6.041
Proposed *3D-CNN*$_{all}$	**3.198**	**3.274**	**3.109**	**3.280**	**3.073**	**3.700**	**2.608**	**2.758**	**3.159**	**3.024**	**3.560**	**3.568**
Proposed *3D-CNN*$_{live\ II}$ *and 3D-CNN*$_{live\ I}$	6.852	6.701	5.893	5.648	7.136	7.303	4.202	**3.735**	5.086	**3.547**	**4.929**	**4.974**

Table 4.4: Asymmetric/symmetric SROCC scores using the 3D LIVE Phase-II dataset.

Method	*Symmetric*	*Asymmetric*
Benoit [12]	0.860	0.671
You [130]	0.914	0.701
Gorley [45]	0.383	0.056
Chen [21]	**0.923**	**0.842**
Hewage [50]	0.656	0.496
Bensalma [13]	0.841	0.721
Akhter [5]	0.420	0.517
Proposed *3D-CNN*$_{all}$	**0.939**	**0.947**
Proposed *3D-CNN*$_{live\ I}$	**0.883**	**0.875**

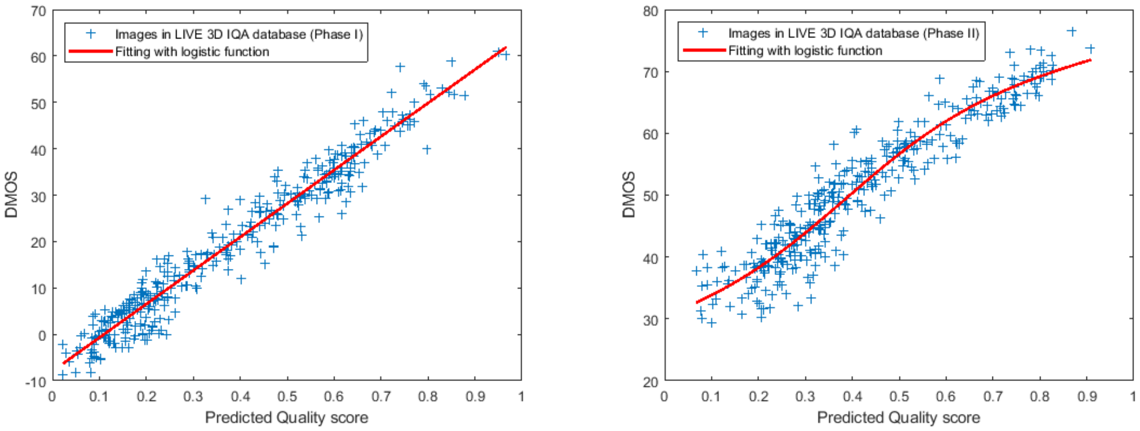


Figure 4.3: Scatter plot of DMOS (subjective scores) versus scores from the proposed metric (*3D-CNN*$_{all}$) on both LIVE 3D Phase I and Phase II databases.

Notice that italicized methods are no-reference metrics, the others are full-reference metrics. The experimental results show that the proposed scheme has good consistency among four distortion types with human subjective evaluation. Since that the 3D-CNN$_{all}$ model has been trained on larger datasets, including the two databases, allowing a better fit model. However, the 3D live phase I database shares four scenes with the phase II database, which might increase the overall model's performance.

4.2.5 Conclusion

A new no-reference IQA metric is proposed for stereoscopic images. That outperforms most of the existing stereo IQA methods. The measure is based on human binocular perception and learning structure evaluation. The handcrafted features have been replaced by deep learning. Four CNN models have been used to map the cyclopean view to quality score. The obtained results indicate that the metric correlates with subjective scores DMOS over symmetric and asymmetric distortions. The performance also indicates that the synthesized Cyclopean image is reliable for perceptual quality evaluation. In the following work, we explore other CNN architectures to assess the stereo-pair image quality.

4.3 2nd model: NR-SIQA based on Deep Features from Cyclopean Image

4.3.1 Approach overview

Most of the proposed SIQA approaches use handcrafted quality features that are derived manually from the stereoscopic picture. With the use of Deep Learning, the suggested approach allows learning quality features from the input data automatically. However, a work has been done in [58] that explore this concept and propose NR SIQA metric. The authors pursued a two steps of training. They first trained the CNN model to extract features from small stereo-pair image patches. The model is then followed by feature concatenation layer and regression layer for second training to predict the quality. Another blind CNN-based metric has been proposed by [125]. The authors have used end-to-end dual stream CNN with multi-level feature concatenation through the network. The proposed metric as shown in Fig. 4.4 involves three simple steps: in the first step, the cyclopean image is computed. Second, divide the cyclopean image into four equivalent parts and train four CNN models that generate a feature bank. Third, the quality score is predicted from the extracted features using a SVR.

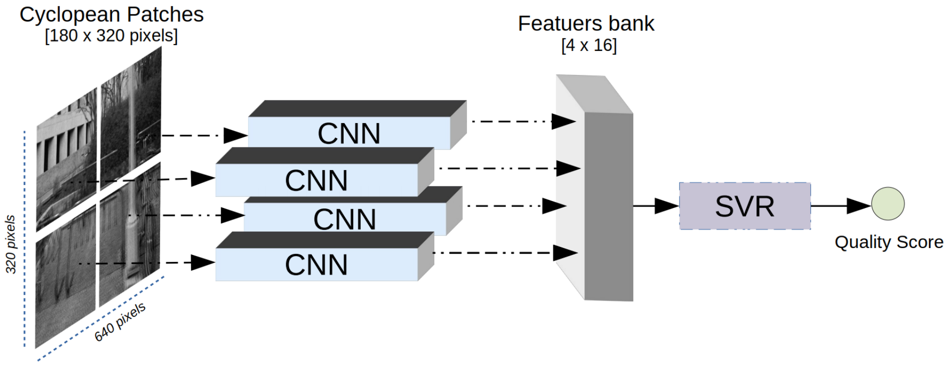


Figure 4.4: Flowchart of the proposed method.

4.3.2 Deep Feature extraction

Generally, at each region corner of the cyclopean image, the structure differs e.g. textures, color and pixel intensities. As we want to derive various quality features, we simply divide the input cyclopean image into four equivalent patches. This partition covers the four corners and deals with different structures individually. Four CNNs are then needed to extract quality feature sets from each structure. In the case of using just one CNN, the model will tend to extract the general characteristics since the network weights remain the same. The four trained CNNs have similar architecture but different weights, thus they enrich the features bank as shown in Fig. 4.4. Meanwhile, we assume that using two models will provide fewer quality indicators than four. For the feature extraction, we design a light-weight CNN model from scratch and compare its performance with most common pre-trained models: AlexNet [62], VGG-16 and VGG-19 [103], Resnet18 [47], Inception-v1 [112].

The cyclopean image is thus fed to the CNN models to to extract quality-aware indicators. Each CNN expects an input of size 180 × 320 pixels. For each patch, one CNN model is trained and used to extract a vector of size 1×16. The suggested CNN architecture consists of 12 layers as shown on Fig. 4.5, after each convolution layer, a batch normalization layer is applied to speed up the learning [54], followed by a ReLU layer as activation function and Max-pooling layer to reduce dimensionality. The network includes three convolutional layers. The first and second convolution layers produce 64 filters of size $[11 \times 11]$. While the third convolution layer has 32 filters of size $[5 \times 5]$. All convolution layers have 4 pixels stride in both horizontal and vertical directions. The first used Max-pooling layer has a size of [7 × 7] and a stride of [2×2] pixels. The second Max-pooling layer has a size of [5 × 5] and a stride of [1×1] pixel. After all, a fully connected 1 ×16 layer is used to provide 16 elements quality indicators. The four networks are trained for 150 epochs with a learning rate of 0.01. Stochastic Gradient Descent (SGD) with momentum has been applied to update the network weights. The extracted feature vectors [4×16] is then fed to a SVR model with a Gaussian kernel function to predict the quality scores.

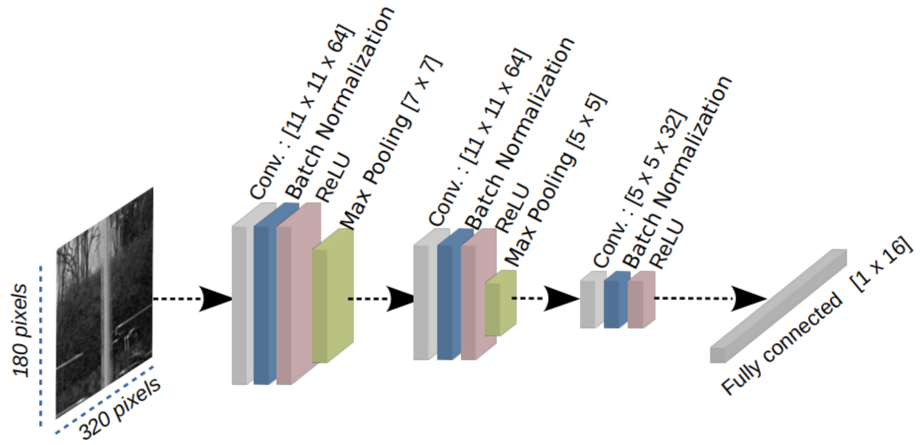


Figure 4.5: The proposed Convolutional Neural Network architecture for feature extraction.

4.3.3 Datasets and training protocol

As used in earlier suggested metrics. The two databases LIVE 3D phase I and phase II were used to test the efficiency of our metric. We normalize the train set outputs (DMOS) to min-max normalization [0 to 1], where the closest to zero the better quality is. The 5-fold cross validation technique has been adopted. The dataset is split into 5 folds, where each fold is divided to 80% train set and 20% test set chosen randomly. The protocol has been repeated for 10 iterations to show the generalization ability of our method and the mean performance values are reported. The performance has been measured across three metrics as in previous metrics: The *RMSE*, the *PLCC*, and *SROCC*.

4.3.4 Experimental results: Quality evaluation

We first evaluated the relevance to use SVR as regressor instead of FC layers, usually done. To this end, the SVR has been replaced by a FC regression layer. Table 4.5 indicates the comparison PLCC correlation results of the designed network model on LIVE 3D II database. The combination of CNN and SVR has increased the quality prediction accuracy compared to CNN model alone. In addition, the use of four different CNN models enrich the quality features bank to improve the overall quality prediction. Although the proposed CNN architecture ends up with PLCC of 0.932 over LIVE-II dataset, we furthermore test and investigate the performance of six common pre-trained CNN models. Where the same training protocol and configurations of our designed model were used. Each pre-trained model has been adjusted and then used to extract [1 x 16] feature vector from each patch.

Table 4.5: PLCC correlation results of our CNN regression model vs. CNN + SVR combination over the four patches from LIVE 3D II database.

Number of patch	1	2	3	4	All
CNN + FC regressor	0.918	0.920	0.922	0.905	0.910
CNN + SVR regressor	0.927	0.928	0.932	0.907	0.932

Table 4.6 presents these experiments using the LIVE-II database and the best ranked extractor was found to be vgg-16. Overall, the pre-trained models except alexnet outperform our CNN design which was expected since the pre-trained CNNs are large and deeper networks. For instance, vgg-16 has about 138 million (approx) parameters while alexnet has around 62 millions. Alexnet gives the lowest correlation performance among all models. The vgg-16 and vgg-19 yield similar correlation performance with little differences since they have nearly the same architecture. These models contain more series of convolutional layers than our architecture and thus extract higher and better quality indicators for prediction. In the meantime, going deeper than vgg-16 model, resnet and inception extractors appear to slightly diverge from the path toward the best indicators. However, our built CNN is almost two times faster than vgg-16. The run-time indicates

108 ms (milliseconds) for vgg-16, and 56 ms for our CNN using the same hardware and stereoscopic image. With the provided competitive performance, this will be beneficial in case of limited resources. Otherwise, the implementation of vgg-16 would be better choice.

Table 4.6: Performance of different pre-trained feature extractors on LIVE-II database.

Model	PLCC	SROCC	RMSE
AlexNet	0.922	0.921	4.355
VGG-16 Gray	**0.948**	**0.941**	**3.817**
VGG-19	0.946	0.938	3.888
Resnet18	0.930	0.930	4.122
Resnet50	0.940	0.939	3.894
Inception-v1	0.938	0.939	3.897

Our method has been then compared with several FR and NR SIQA metrics, including six FR and seven NR SIQA metrics. Tables 4.7, 4.8 and 4.9 show the results of all SIQA algorithms on LIVE 3D phase I and phase II databases. The best outcome of NR category is highlighted in bold. We reported the outcomes of using the scratched CNN and the pre-trained vgg-16. The results obtained on LIVE 3D Phase I show the efficiency of our method, since it outperforms all the compared metrics in terms of SROCC and RMSE, including FR ones. For FF distortions, Karimi *et al.* [57] metric obtained better results, but our method remains the best on the rest type of distortions. Meanwhile in term of PLCC, our method has the best correlation on the five distortions.

Table 4.7: SROCC results on the 3D LIVE Phase-I and Phase-II databases.

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	0.923	0.751	0.867	0.455	0.773	0.728	0.923	0.751	0.867	0.455	0.773	0.728
You [130]		0.909	0.894	0.795	0.813	0.891	0.786	0.909	0.894	0.795	0.813	0.891	0.786
Gorley [45]		0.875	0.110	0.027	0.770	0.601	0.146	0.875	0.110	0.027	0.770	0.601	0.146
Chen [21]		0.940	0.814	0.843	0.908	0.884	0.889	0.940	0.814	0.843	0.908	0.884	0.889
Hewage [50]		0.880	0.598	0.736	0.028	0.684	0.501	0.880	0.598	0.736	0.028	0.684	0.501
Bensalma [13]		0.905	0.817	0.328	0.915	0.915	0.874	0.938	0.803	0.846	0.846	0.846	0.751
Akhter [5]	NR	0.714	0.724	0.649	0.682	0.559	0.543	0.714	0.724	0.649	0.682	0.559	0.543
Zhou [135]		0.921	0.856	0.562	0.897	0.771	0.901	0.936	0.647	0.737	0.911	0.798	0.819
Fang [35]		0.883	0.880	0.523	0.523	0.650	0.877	0.955	0.714	0.709	0.807	0.872	0.838
Chen [20]		0.926	0.839	0.832	0.951	0.918	0.920	0.910	0.825	0.843	0.929	0.896	0.852
Kim [58]		-	-	-	-	-	-	0.922	0.885	0.763	0.932	**0.945**	0.938
Karimi [57]		0.945	0.917	0.750	0.919	**0.837**	0.947	0.953	0.875	0.832	0.874	0.907	0.913
Liu [69]		0.951	0.888	0.785	0.917	0.821	0.928	0.946	**0.909**	0.825	**0.936**	0.938	0.901
Proposed Gray		0.925	0.921	0.666	0.924	0.799	0.928	0.928	0.897	0.809	0.900	0.880	0.909
Proposed vgg-16 Gray		**0.964**	**0.943**	**0.834**	**0.953**	0.803	**0.956**	**0.959**	0.888	**0.875**	0.935	**0.945**	**0.948**

Table 4.8: PLCC results on the 3D LIVE Phase-I and Phase-II databases.

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	0.926	0.784	0.853	0.535	0.807	0.784	0.926	0.784	0.853	0.535	0.807	0.784
You [130]		0.912	0.905	0.830	0.784	0.915	0.800	0.912	0.905	0.830	0.784	0.915	0.800
Gorley [45]		0.796	0.485	0.312	0.852	0.364	0.451	0.322	0.372	0.874	0.934	0.706	0.515
Chen [21]		0.957	0.834	0.862	0.963	0.901	0.907	0.957	0.834	0.862	0.963	0.901	0.907
Hewage [50]		0.891	0.664	0.734	0.450	0.746	0.558	0.891	0.664	0.734	0.450	0.746	0.558
Bensalma [13]		0.914	0.838	0.838	0.838	0.733	0.887	0.943	0.666	0.857	0.907	0.909	0.769
Akhter [5]	NR	0.772	0.776	0.786	0.795	0.674	0.568	0.929	0.772	0.776	0.786	0.795	0.674
Zhou [135]		-	-	-	-	-	0.929	-	-	-	-	-	0.856
Fang [35]		0.900	0.911	0.547	0.903	0.718	0.880	0.961	0.740	0.764	0.968	0.867	0.860
Chen [20]		-	-	-	-	-	0.937	-	-	-	-	-	0.937
Kim [58]		-	-	-	-	-	-	0.910	0.910	0.768	0.951	0.957	**0.941**
Karimi [57]		0.955	0.939	0.771	0.959	0.882	**0.956**	0.966	0.897	0.866	0.957	0.918	0.923
Liu [69]		0.966	0.938	0.810	0.956	0.855	0.945	**0.969**	0.936	0.867	**0.987**	**0.959**	0.913
Proposed Gray		0.936	0.905	0.811	**0.967**	**0.887**	0.911	0.931	**0.944**	0.689	0.951	0.851	0.932
Proposed vgg-16 Gray		**0.970**	**0.960**	**0.845**	0.962	0.865	0.955	0.959	0.887	**0.888**	0.981	0.931	**0.941**

Table 4.9: RMSE results on the 3D LIVE Phase-I and Phase-II databases.

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]		4.028	6.096	3.787	11.763	6.894	7.490	4.028	6.096	3.787	11.763	6.894	7.490
You [130]		4.396	4.186	4.086	8.649	4.649	6.772	4.396	4.186	4.086	8.649	4.649	6.772
Gorley [45]		5.202	9.113	6.940	4.988	8.155	9.675	5.202	9.113	6.940	4.988	8.155	9.675
Chen [21]	FR	3.368	5.562	3.865	3.747	4.966	4.987	3.368	5.562	3.865	3.747	4.966	4.987
Hewage [50]		10.713	7.343	4.976	12.436	7.667	9.364	10.713	7.343	4.976	12.436	7.667	9.364
Bensalma [13]		-	-	-	-	-	7.558	-	-	-	-	-	7.203
Akhter [5]		7.416	6.189	4.535	8.450	8.505	9.294	7.416	6.189	4.535	8.450	8.505	9.294
Zhou [135]	NR	-	-	-	-	-	6.010	-	-	-	-	-	6.041
Fang [35]		-	-	-	-	-	7.191	-	-	-	-	-	5.767
Karimi [57]		5.017	4.644	4.290	4.458	**5.997**	4.998	**2.936**	5.083	4.071	4.581	4.974	4.436
Liu [69]		-	-	-	-	-	5.268	-	-	-	-	-	7.658
Proposed Gray		6.046	4.246	4.725	4.419	6.502	5.905	3.770	**4.160**	4.280	3.506	5.288	4.629
Proposed vgg-16 Gray		**4.013**	**3.597**	**3.488**	**3.943**	6.223	**4.865**	3.002	4.526	**3.366**	**2.685**	**4.176**	**3.817**

For LIVE 3D phase II, the same behaviour has been noticed with the best overall performance and competitive results on the degradation types. Compared to the results on LIVE 3D Phase I, we obtained higher Spearman's correlations for WN, JPEG and FF, but still not high as other degradation types.

Table 4.10 shows the performance on symmetric and asymmetric distorted stimuli. As can be seen, performances of all method are often higher for symmetric distribution. Our method outperforms most of the compared FR and NR metrics with 0.936 and 0.953 as SROCC for symmetric and asymmetric distributions, respectively. Hence, the proposed scheme has good correlation with human subjective evaluation across four types of distortion as well as symmetric/asymmetric distributions.

Table 4.10: Asymmetric versus Symmetric SROCC results on 3D LIVE Phase II database.

Distortion Type	Benoit [12]	You [130]	Gorley [45]	Chen [21]	Hewage [50]	Bensalma [13]	*Akhter* [5]	*Proposed*	*Proposed vgg-16*
Symmetric	0.860	0.914	0.383	0.923	0.656	0.841	0.420	0.921	**0.936**
Asymmetric	0.671	0.701	0.056	0.842	0.496	0.721	0.517	0.909	**0.953**

Scatter plots that exhibit the prediction responses against human score (DMOS) on LIVE

3D phase I and phase II are given in Fig. 4.6. As can be seen, the distribution of the predicted scores well fit the DMOS with low dispersion. According to the different degrees of deformations/distortions. Each distortion type scores are well spread according to human predictions. This can show a consistency performance for all types distortion.

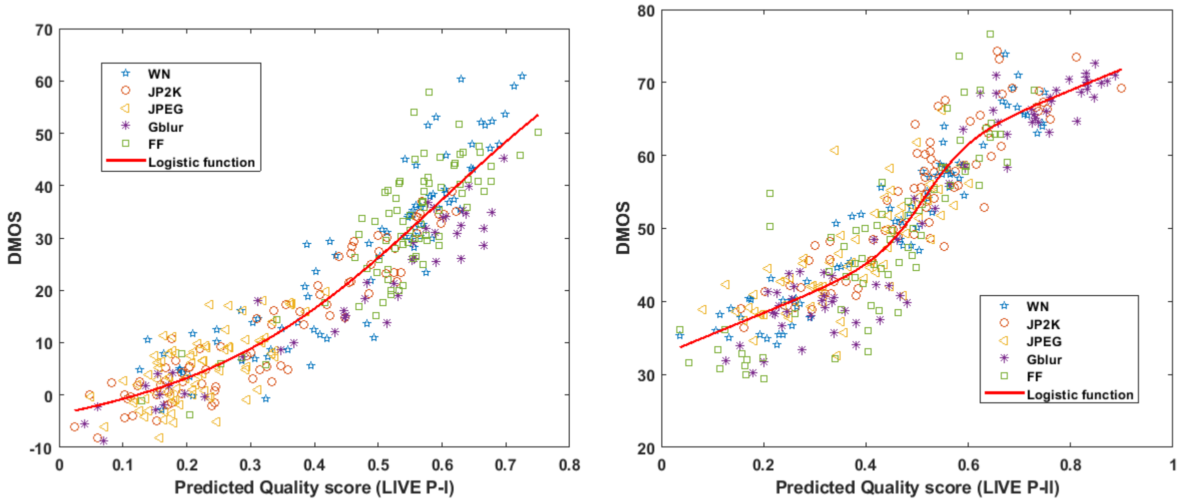


Figure 4.6: Scatter plot of subjective scores DMOS against scores from the proposed metric using the designed CNN on LIVE 3D Phase I and Phase II databases.

4.3.5 Quality indicators visualization

In this section, we investigate the extracted features by the designed networks (shown in Fig. 4.5), and examine which parts of the cyclopean image are most important for our CNN models. A patch was chosen from the cyclopean which formed using distorted stereo images. These latter are fed to a trained CNN model as test patches and then inspect the outputs of activation functions (ReLU) after the first and second convolutional layers. The convolution layer produces 64 channels. Among the 64 channels output from ReLU layer, their mean values are computed and the strongest channel has been selected by indexing the maximum. Fig. 4.7 despite the first and second ReLU layer responses for the input cyclopean patch that were constructed under three types of distortions: WN, JPEG, and Blur. As can be seen, where the warmer (closer to 1) regions activate the ReLU function and thus influence the decision of the network. It is remarkable that the first activation function reflect the presence of pixel deformation. The JPEG compression is well known artifact that causes undesirable blocks in the image due to the quantization.

This issue is stated in ReLU 1 activation map of JPEG patch that shows the selection of these blocks as a highly important information to pass through the network. As well as for WN and Blur cyclopean patches, the ReLU 1 activation function have succeeded to focus on noise and blur artifacts. However, additionally, with the help of this activation function, we can form a distortion map. The latter can then be used by enhancement algorithms to concentrate on the most damaged regions instead of analyzing the while scene.

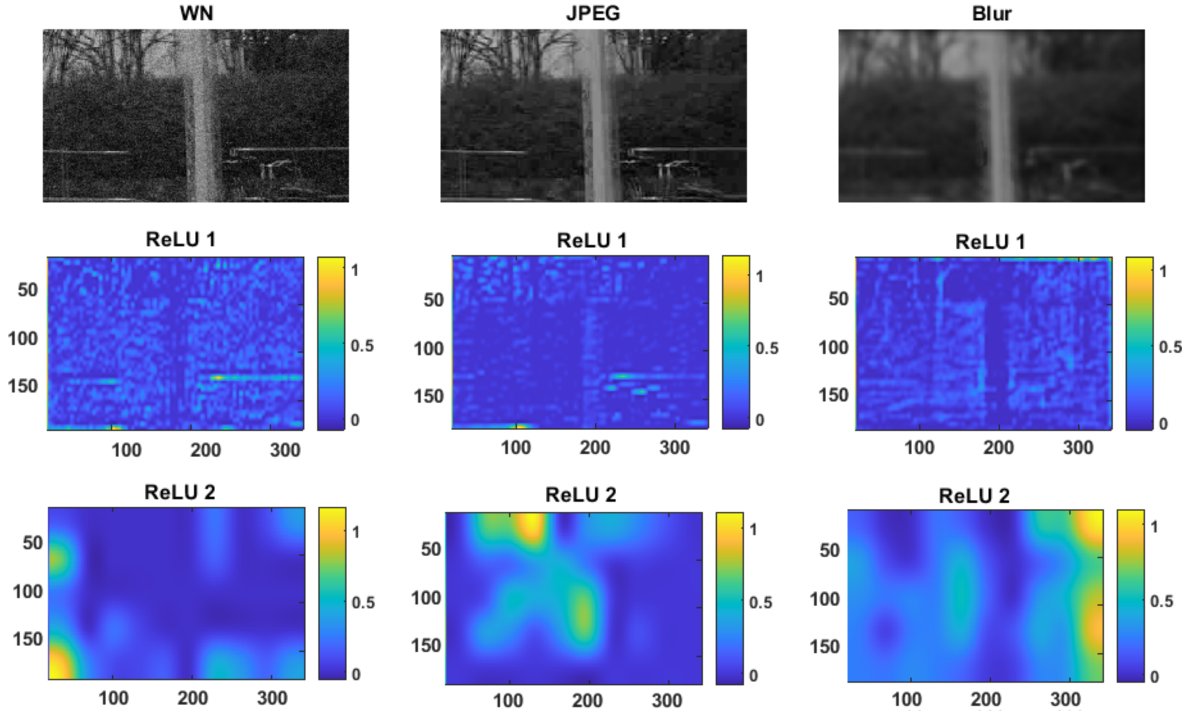


Figure 4.7: The first and second ReLU activation layer outputs from a test cyclopean patch for three degradation types.

While the second activation function (ReLU layer) is controlled by a deeper representation that makes it harder to fully comprehend the outputs. However, for JPEG cyclopean patch, most deformed regions are placed above and by the edge of a pillar in the scene. Meanwhile for Blur, the deformed regions are located around everywhere the pillar. From the second ReLU output maps, the warmer regions are somewhat distributed according to the most infected regions in the scene. For further analysis, Fig. 4.8 provides visualisation of the extracted features from each patch. For comparison, three of the same scene cyclopean images of different distortion types and degrees were used. A quality score has been computed via the proposed scheme for each patch. As can be seen, the feature values are within range of 0 to 1 appear diversity as the degree of degradation varies. Note that the blue dots refer to features from non distorted stereoscopic image input. The distribution of blue dots are similar in all patches. The orange and red feature distributions refer to distorted stereoscopic image inputs. Here we notice non similar distribution at each patch because the approach tends to extract quality features relevant to the spatial information at each corner of the scene (as discussed earlier in section deep feature extraction). Consequently, each model derives distinct features and enrich the feature bank which is utilized to measure the quality. With regard to these observations, we can conclude that the trained networks focus on the pixel deformations to extract a complex quality indicators. The decision that defines these indicators is then guided by the type and degree of distortions.

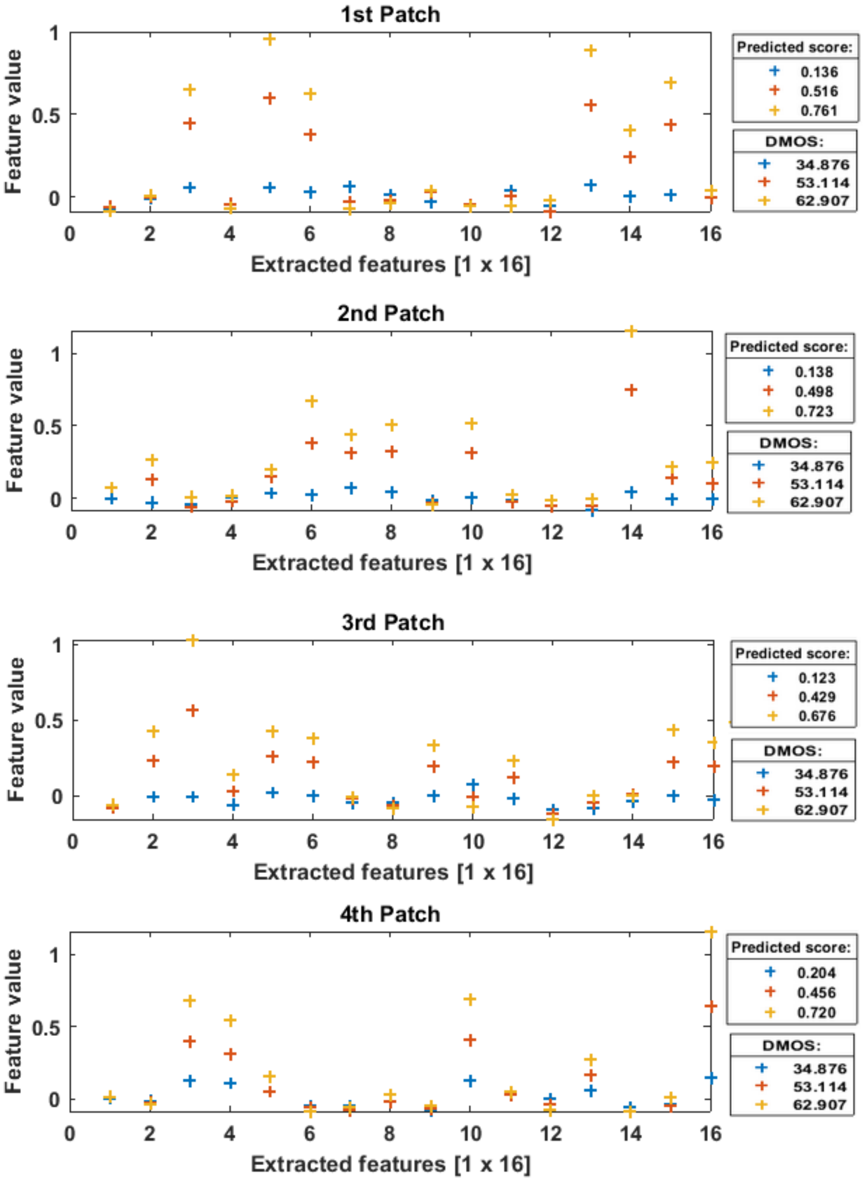


Figure 4.8: Extracted features bank from three cyclopean images of the same scene. Each plot represents the sixteen extracted quality indicators from a different patch. The first to the fourth patch from above to below, respectively.

4.3.6 Conclusion

In this work, a new deep feature extraction approach has been explored for NR SIQA. The simplicity of proposed scheme is an advantage for implementation in the multimedia software. As in previous work, the proposed metric uses cyclopean image hypothesis that considers binocular rivalry phenomenon. Then, four CNN models are used to extract bank of features from the cyclopean image. This bank is then mapped to a quality score using a SVR. The obtained results have corroborated the correspondence between the proposed metric and the subjective DMOS over asymmetric and symmetric distributions. Based on the performance achieved, the followed workflow that combines multi-extractors with SVR could be useful for future works.

Chapter 5

Contributions based on HVS modeling and Saliency information

5.1 Introduction

In the majority of NR-SIQA model designs, quality indicators of image structure, play essential roles as discussed in chapter 2. However, the distortions added to images generate changes in structural features which can be captured by structural feature statistics. Based on how these quality-aware features are calculated, NR models can be further categorized into machine learning-based methods and training-free based methods. Training-free approaches have an internal generalization potential, and yet, their performances are currently inferior to machine learning-based methods. Instead, using machine learning techniques such as SVR and Random Forest (RF), image feature values can be simply mapped to the image quality index, assisting machine learning-based NR-IQA models to obtain comparatively higher evaluation performance. Furthermore in latest years, deep-learning-based algorithms that directly map an image or image structure to a quality index have achieved promising results. But, there are several flaws to this latter, such as fixed input pixel resolution, pixel attack sensitivity, and large scale training data requirement. Regardless of the learning approach employed in the SIQA system, modeling HVS is important to simulate the visual judgment. However, because the HVS is a complex visual process and still an open question for researchers. In SIQA design, many researchers have used fusion hypothesizes of the perceived left and right eye signals called cyclopean view [21, 13]. Meanwhile, in most of the suggested SIQA approaches the human visual attention is not explored.

This chapter explores whether visual attention should be taken into account when developing an objective SIQA metric. To that end, we present a new metric that takes saliency information into account [76]. Several tests, including an ablation test, are used to verify the results of this experiment. Furthermore, we investigate the effect of distortions on the 3D saliency map.

5.2 NR-SIQA using Deep Quality evaluator guided by 3D Saliency

5.2.1 Approach overview

The general framework of the proposed method is summarized in Fig. 5.1. From a given stereo image, the cyclopean image is first calculated, allowing to consider the binocular rivalry phenomenon as mentioned above [15]. Then, the 3D saliency map of the stereo image is computed. It aims to focus on regions that attract more our perception. After having thresholded the obtained 3D saliency map, small patches are extracted and fed into a CNN model in order to predict the overall quality of the stereo image. Each of these steps is described in coming subsection. As seen in previous works, the disparity map is here computed using an SSIM-based method (discussed in Chapter 3).

5.2.2 RGB Cyclopean image

Followed by the study conducted in Chapter 3 that exhibits the benefit of using cyclopean image for SIQA. Where the use and non-use of cyclopean hypothesis has been analyzed. The comparison results indicated better accuracy when cyclopean image is being deployed. Inspired by the model used in previous work, we construct a cyclopean image over three

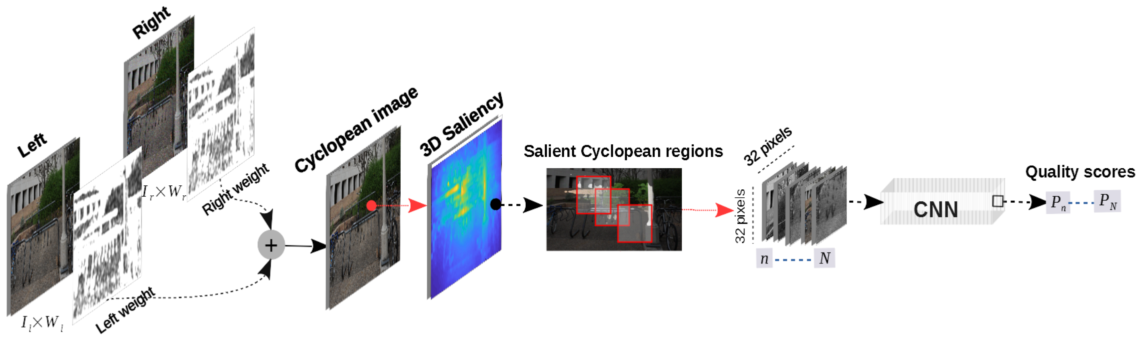


Figure 5.1: Flowchart of the proposed metric.

channels Red, Green, and Blue (RGB) rather than one gray channel to maintain the distortion effects on the stereo image. The formula used is as follows:

$$C(x, y)_n = w_l(x, y)_n \times I_l(x, y)_n + w_r(x + d, y)_n \times I_r(x + d, y)_n \tag{5.1}$$

where C refers to the cyclopean image and n for the color channel number in-which $n \in \{R, G, B\}$. Left and right views are represented by I_l and I_r, respectively. d is the disparity index that matches pixels from left image I_l with those in right image I_r. While w_l and w_r are the weights of the left and right eyes, respectively. The weights w_l and w_r are defined in 3.4 and 3.5, respectively.

As an example presented in Fig. 5.2, sub-Fig. (a) shows RGB cyclopean image formed from the left image in sub-fig (b) that is not distorted and the right image in sub-fig (c) that suffers from WN distortion. It is worth noticed that this asymmetric distortion is stated clearly onto the cyclopean image (the red boxes).

5.2.3 3D Saliency map

Visual attention/saliency is an important characteristic of our HVS since it represents the regions of the image in which the observer focus the most. Hence, salient regions impact more the subjective scores given by the observers and thus the quality of a given image is highly related to these regions. While the use of the 3D saliency map is a further move in HVS simulation, it remains unconsidered in the most current NR-SIQA metrics.

According to this observation, 3D saliency map has been used in this study to extract perceptual relevant patches instead of all patches. The 3D saliency method suggested in [115] has been here used. This method is based on the integration of the depth information and 2D saliency maps. The saliency map of the luminance, color and texture from one view are first computed [36]. Then, the depth map is calculated through the left and right views as shown in Fig.5.2.d. After a normalization step, the 3D saliency map is finally given by averaging the achieved maps. For comparison example, we compute non-depth saliency map (i.e. 2D saliency) and depth saliency map (i.e. 3D saliency) displayed in Fig. 5.2.e and Fig. 5.2.f, respectively. We can see that the 3D saliency map gives more importance to near objects than the 2D one because the algorithm incorporates the depth map.

The 3D saliency map is then normalized (using min-max normalization) and thresholded to extract patches of size 32x32x3 from the cyclopean image allowing thus to focus only on the most salient regions. The extracted patches are then fed to a CNN model to predict the quality. After several tests, the threshold has been fixed to 0.3. The impact of the threshold value on the performance is presented in Section 5.2.6.

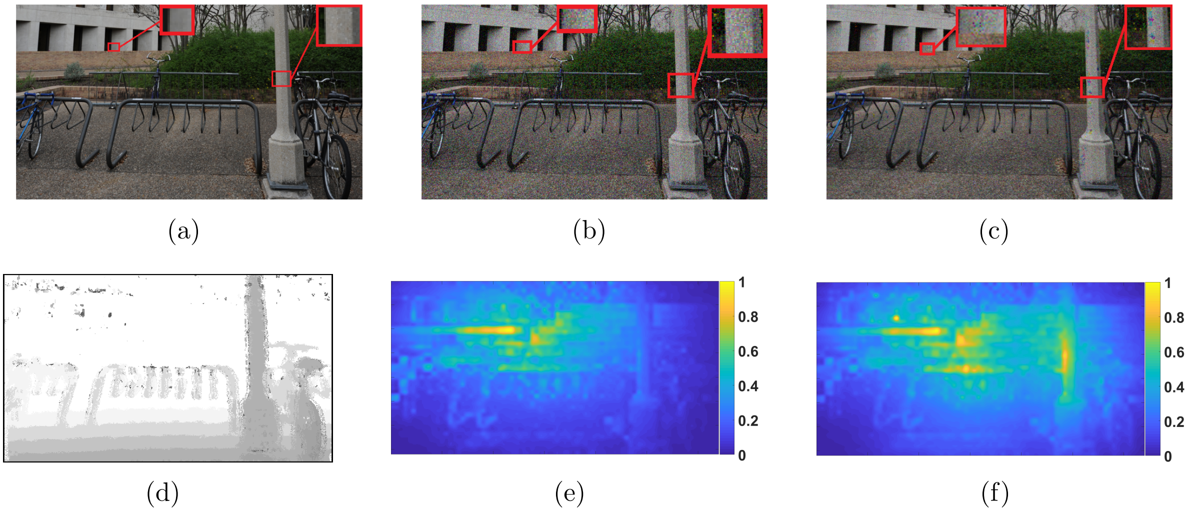


Figure 5.2: Saliency of a stereo image: (a) Left view without distortion, (b) Right view with White Noise (WN) distortion, (c) Synthesized RGB cyclopean image, (d) estimated depth using disparity map, (e) 2D saliency map and (f) the used 3D saliency map.

5.2.4 Quality prediction model

As discussed earlier (in Chapter 4), several CNN models with different architectures have been proposed in the literature. In this work, performances of five pre-trained models widely used have been compared, briefly described above:

- **AlexNet [62]:** Developed in 2012, the AlexNet model is one of the pioneering models proposed by Alex Krizhevsky. This model highlighted the relevance of using CNN models for classification tasks. Composed of 5 convolutional layers and 3 FC layers, the authors stressed three main points: the use of the Relu (Rectified Linear Units) function, the exploitation of the dropout to prevent the over-fitting and overlap during the pooling step.

- **VGG16 and VGG19 [103]:** have been proposed in 2014. VGG models were developed by the Oxford Visual Geometry Group. To increase the ability of the model to discriminate between objects, the authors integrated more non-linearities by using convolutional layers with 3x3 filters instead of 7x7 filters. Several versions were proposed with 11 (VGG11), 13 (VGG13), 16 (VGG16) and 19 (VGG19) layers. Here, VGG16 and VGG19 are used and compared.

- **ResNet18 and ResNet50** [47]: In 2015, a Residual Neural Network (ResNet) model was proposed. This model stands out by its integration of a residual module. The idea developed by the authors is to reformulate the output (H(x)=F(x)) of each series of Conv-ReLu-Conv by adding the input $\boldsymbol{x}$ as information ($\boldsymbol{H(x) = F(x){+}x}$). Different versions are available: ResNet18 (18 layers), ResNet34 (34 layers), ResNet50 (50 layers), ResNet152 (152 layers) and so on. ResNet18 and ResNet50 are used in this study.

The use of these models allows to compare different depths (from a shallow model i.e. AlexNet to deeper models i.e. the other models), different architectures (ResNet and VGG) as well as same architecture with different depths (VGG16 against VGG19 and ResNet18 against ResNet50).

Each of these models has its specificities as shown in Table 5.1 that compares the used pre-trained models in terms of memory size and amount of learned parameters. The network depth refers to the largest number of sequential convolution or fully connected layers on the path from the input layer to the output layer. They have a distinct number of learnable parameters and different depth sizes. This diversity will drive us to the best architectures that are suited for quality assessment. It is worth noticed that these models were modified and fine-tuned to adapt their learnable parameters to our context.

Table 5.1: Pre-trained models descriptions.

Model	Size	Learnable parameters (Millions)	Depth
AlexNet	227 MB	61.0	8
VGG-16	528 MB	138	16
VGG-19	549 MB	144	19
ResNet18	44 MB	11.7	18
ResNet50	98 MB	25.6	50

During the learning, each pre-trained CNN model is fine-tuned for 50 epochs using a learning rate of 0.01. Stochastic Gradient Descent (SGD) with a momentum equals to 0.9 is used as optimization function. The human scores are normalized in the form of DMOS/MOS to min-max normalization [0,1]. The closer to 0 the better quality of the stereo image is for DMOS and the opposite for MOS. After-all, the quality index of a given stereo image is computed by averaging the predicted scores over the extracted saliency patches.

5.2.5 Datasets and Training Protocol

To examine the consistency and effectiveness of our method, four databases have been used to evaluate the performance of our metric. These datasets are listed briefly in Table 5.2.

Table 5.2: Summary of the four databases.

Database	# of Reference scenes	Resolution	# of images (Sym., Asym.)	Distortions
3D LIVE P-I	20	360 x 640	365 (365, 0)	JP2K, JPEG, WN, Blur, FF
3D LIVE P-II	8	360 x 640	360 (120, 240)	JP2K, JPEG, WN, Blur, FF
Waterloo IVC 3D P-I	6	1080 x 1920	330 (180, 150)	JPEG, WN, Blur
Waterloo IVC 3D P-II	10	1080 x 1920	460 (210, 250)	JPEG, WN, Blur

It is worth to remember that the asymmetric degradations in the Waterloo P-1 and P-2 databases are different from those in the LIVE-II database. LIVE-II uses only one type of distortion to perform the asymmetry, while the two Waterloo databases consider the possibility of multiple types of degradation in which the left and the right images are affected by different distortions.

Generally, the above-described SIQA databases have small-limited labelled images. To increase the amount of data, data augmentation is often applied. The available data augmentation techniques except horizontal flipping, affects the subjective quality ratings. The rotation and re-sizing approaches often applied change the observers perception of spatial details and are thus not appropriate for SIQA methods. Therefore in this work, neither rotation nor translation or re-sizing were applied. Instead of, we allow a maximum of 80% overlapping between patches. The expected quality rating for each scene is the average of quality scores obtained from patches, described as follows:

$$Q = \frac{1}{N} \sum_{n=1}^{N} P_n \tag{5.2}$$

Where P_n is the predicted score for the n^{th} patch, N is the number of patches, and Q is the final quality score. We have carried out 10-fold validation test by randomly splitting the dataset into training (80%) and test (20%) at each time. The average result is then used as evaluation criterion. We also evaluate the generalization ability of our method by applying a cross-dataset evaluation. The performance has been measured across the usual three indexes: The *RMSE*, *SROCC*, and the *PLCC*. Note that objective scores are fitted to the subjective ones using logistic function [100].

5.2.6 Different Saliency thresholds and predictors analysis

In this section, many tests have been conducted to define the best network architecture and to identify suitable saliency threshold. Saliency-based patches are extracted with regard to threshold value. The five pre-trained models are adjusted and tested using the same train configurations as discussed in section 5.2.4. Starting with value of 0.1, we update the threshold and notice the performance using the LIVE P-2 database in Table 5.3. The Table also includes the number of patches extracted at each threshold.

Table 5.3: PLCC results of different deep models versus saliency threshold on LIVE-P2.

Saliency threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Number of patches	82277	68900	46300	23968	12276	5466	1692	557	369
AlexNet	0.960	0.968	0.970	0.969	0.959	0.906	0.870	0.832	0.801
VGG-16	**0.977**	**0.984**	**0.985**	**0.983**	**0.981**	**0.945**	**0.907**	**0.891**	**0.881**
VGG-19	**0.977**	0.983	0.984	0.982	0.980	0.943	**0.907**	0.890	**0.881**
ResNet18	0.970	0.976	0.975	0.974	0.968	0.926	0.889	0.794	0.500
ResNet50	0.966	0.974	0.976	0.975	0.972	0.931	0.882	0.823	0.675

Obtained PLCC results show that the VGG-16 and VGG-19 architectures are better for mapping the extracted patches to quality scores. From plots in Fig. 5.3, we notice that using different saliency-based cropping thresholds influence the quality prediction with best threshold value of 0.3. As we increase the starting value, we get better results for all models. After threshold of 0.3, the coefficient correlations decrease while saliency thresholds cropping increase. The fact that higher saliency threshold gives smaller datasets, it may limit the model to learn best quality prediction from the salient regions. For instance, 0.3 gives 46 300 patches for training, while only 12 276 patches for 0.5. This is a trade-off between the saliency threshold and the training dataset size that need to be balanced. For example, although using a threshold of 0.1 that yields more training sets (i.e. 82277 patches), the better precision results are still obtained with a threshold equal to 0.3. Based on these results, the saliency-guided cropping step allows to considerably improve the performance. Notice that the performance drops for thresholds which offer small train datasets, such as the 0.6 threshold.

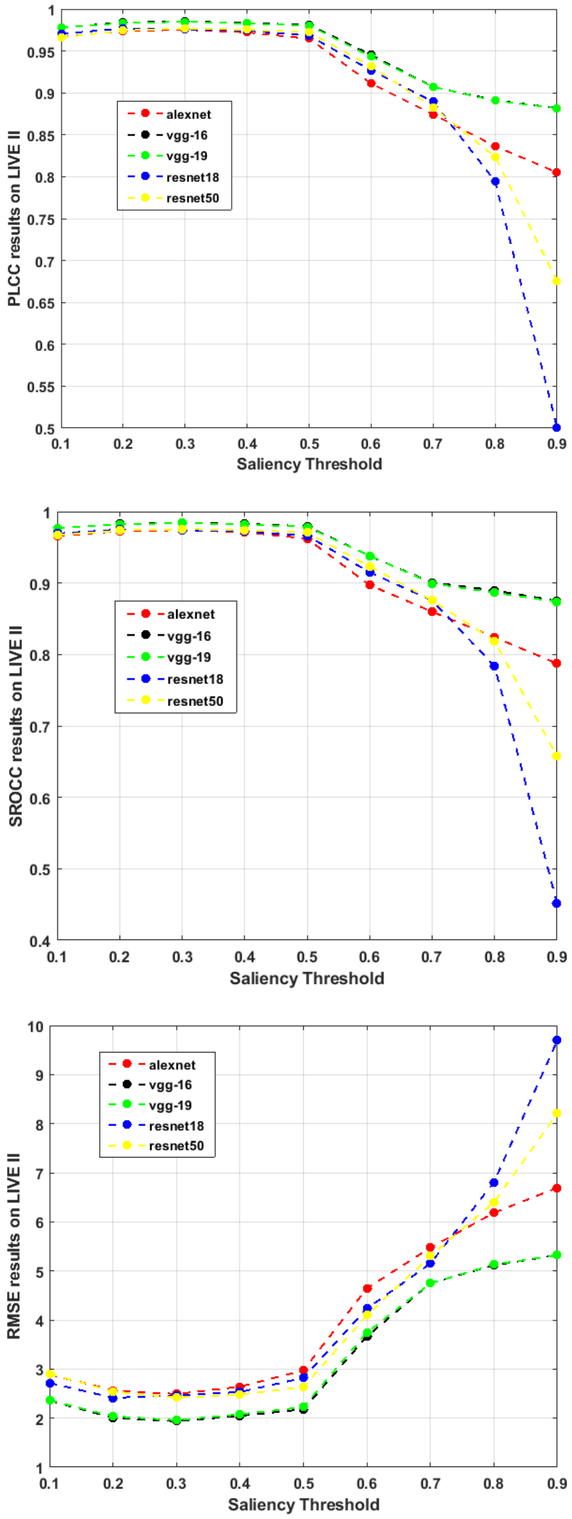


Figure 5.3: PLCC, SROCC and RSME comparison results of pre-trained models versus different thresholds on LIVE-P2.

Moreover, AlexNet gives the lowest correlation performance among all models. VGG-16 and VGG-19 yield similar correlation performance with little differences since they have nearly the same architecture. These models contain more series of convolutional layers and thus extract higher and better quality indicators for prediction. In the meantime, going deeper than VGG-16 model, ResNet18 and ResNet50 regressors appear to slightly diverge from the path toward the best quality predictions. For instance, using the best saliency threshold of value 0.3, AlexNet model with performance of RMSE = 2.491 comes in the last place compared to the the other four networks.

VGG-16 and ResNet18 behave slightly better compared to deeper ones; VGG-19 and ResNet50, respectively. The RMSE is 1.938 for VGG-16 and 2.416 for ResNet18, while the error values for VGG-19 and ResNet50 are 1.957 and 2.459, respectively. Meanwhile, analyzing the same architectures and different depths, VGG-16 performs better than VGG-19. Also ResNet18 provides better results than ResNet50. Despite that going deeper with convolutions improves the accuracy in object recognition/classification tasks, for regression problems it might not perform well. Allowing the network to perform more convolutions does not necessary imply extraction of more precision quality-features.

After the selection of the best pre-trained model and saliency threshold, we evaluated the impact of the saliency-based patch selection and the RGB cyclopean image. For the no saliency test, all possible patches of the cyclopean image were sequentially extracted by sliding over the whole scene from left to right with a stride of 32 pixel (i.e. without overlap). This creates 220 patches for every scene in the LIVE P-2 database, while 128 patches are approximately cropped for the saliency-guided extraction. Table 5.4 shows PLCC, SROCC and RSME results over LIVE P-2 database with and without saliency-guided patches as well as the grayscale cyclopean versus RGB cyclopean as inputs. As can be seen, the saliency-guided patch selection considerably improves the performance with a quality prediction error decrease of 49% in term of RMSE. The use of RGB cyclopean image allows also to increase quality prediction efficiency in both cases (i.e. with and without the saliency-guided patch selection). During subjective assessments, the ratings are given based on RGB stimulus. The RGB cyclopean is therefore closer to reflect the distorted spatial information experienced by the observer. The best result is reached when both are considered. This experiment supports the use of the saliency map and the RGB cyclopean image for the SIQA. Moreover besides accuracy improvement, the saliency guidance approach may also decrease the cost and run-time, since the approach uses recommended patches rather than using all patches of the scene.

Table 5.4: Impact of the saliency-guided patch selection and the RGB cyclopean image on the performance using VGG-16 and a saliency threshold of 0.3. The tests were carried-out on LIVE-P2 dataset.

		LIVE-P2		
Method	Input Stereoscopic image	SROCC	PLCC	RMSE
Saliency guided	RGB	**0.984**	**0.985**	**1.938**
	Gray	0.953	0.960	3.829
Without saliency	RGB	0.958	0.961	3.814
	Gray	0.931	0.942	4.011

5.2.7 Patch-size effect on quality evaluation

The metric implementation needs a fixed size patch for the deep CNN regression stage. The stereo images have different aspects and resolutions. Such change would have an impact on the salient selection regions, and the 32 x 32 patch might not be ideal in this situation. In particular, LIVE-P1 and P2 stereo images have 360 x 640 pixels size, while Waterloo-P1 and P2 stereo images have higher resolution with size of 1920 x 1080 pixels. Tests have been conducted for this manner using the VGG-16 and 0.3 saliency threshold for their best fit. We increase the patch size by 32 x 32 pixels each time and notice the effect on quality prediction performance using the three indexes; PLCC, SROCC, and RSME. Table 5.5 show the results of these tests.

Table 5.5: Performance versus patch size on Waterloo-P1 and LIVE-P2 databases.

	Waterloo-P1			LIVE-P2		
Patch size (in pixels)	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
32 x 32	0.946	0.960	4.376	**0.984**	**0.985**	**1.938**
64 x 64	**0.967**	**0.973**	**3.592**	0.975	0.978	2.342
96 x 96	0.964	0.971	3.757	0.975	0.977	2.389
128 x 128	0.956	0.967	4.009	0.969	0.972	2.628

Performance results demonstrate that increasing patch size can improve the performance for higher resolution stereo images such in Waterloo-P1 and P2 databases. The best patch size for LIVE-P2 is 32 x 32 and 64 x 64 for Waterloo-P1. Typically higher resolution images give the viewer a larger salient region, and increase the number of extracted patches for 32 x 32 pixels cropping.

The number of patches extracted must be balanced by the resolution of the stereo image. Therefore, the patches size relies on the resolution of salient region seen by the observer.

5.2.8 Comparison with the State-of-the-Art

Obtained results have been compared with several FR, RR and NR SIQA. Among them, there are recent blind metrics based on the use of CNN models, namely PAD-Net [128], Chen [23] and Sun [110].

Table 5.6 shows the results of these methods on both LIVE-P1 and P2 datasets. Best metric of each category (FR, RR and NR) is represented on bold and the best one whatever the category is with a gray background. As can be seen, our metric outperforms all the compared NR metrics on both databases. The best FR metric is the one proposed by Chen et al [21], while the method proposed by Ma et al [73] achieved the best performance among the compared RR methods. On LIVE-P1, compared to the best metrics in each category (i.e. Chen for FR and Ma for RR) the improvements in term of PLCC are 7% for FR and 5.6% for RR. While on LIVE-P2, the improvements are 8.8% for FR and 6.3% for RR.

Table 5.6: Overall performance comparison on LIVE-P1 and LIVE-P2.

		LIVE-P1			LIVE-P2		
Type	Metrics	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
FR	Benoit [12]	0.899	0.902	7.061	0.728	0.748	7.490
	You [130]	0.878	0.881	7.746	0.786	0.800	6.772
	Gorley [45]	0.142	0.451	14.635	0.146	0.515	9.675
	Chen [21]	**0.916**	**0.917**	**6.533**	**0.889**	**0.900**	**4.987**
	Hewage [50]	0.501	0.558	9.364	0.501	0.558	9.364
	Bensalma [13]	0.874	0.887	7.558	7.558	0.769	7.203
RR	RR-BPI [89]	-	-	-	0.867	0.915	4.409
	RR-RDCT [74]	0.905	0.906	6.954	0.809	0.843	6.069
	Ma [73]	**0.929**	**0.930**	**6.024**	**0.918**	**0.921**	**4.390**
NR	Akhter [5]	0.383	0.626	14.827	0.543	0.568	9.294
	Zhou [135]	0.901	0.929	6.010	0.819	0.856	6.041
	Fang [35]	0.877	0.880	7.191	0.838	0.860	5.767
	DNR-S3DIQE [87]	0.935	0.943	-	0.871	0.863	-
	Fezza [38]	-	-	-	0.925	0.908	3.018
	3D-AdaBoost [81]	0.930	0.939	5.605	0.913	0.922	4.352
	DBN [128]	0.944	0.956	4.917	0.921	0.934	4.005
	Chen [23]	0.943	0.959	4.838	0.922	0.936	3.667
	Sun [110]	0.959	0.951	4.573	0.918	0.938	3.809
	DECOSINE [127]	0.953	0.962	-	0.941	0.950	-
	StereoQA-Net [134]	0.965	0.973	4.711	0.947	0.957	3.270
	PAD-Net [124]	0.973	0.975	3.514	0.967	0.975	2.446
Gray	Proposed	**0.981**	**0.982**	**3.086**	**0.984**	**0.985**	**1.938**

Overall, the performance of our method from both LIVE datasets is somewhat equivalent with slight advantage for LIVE-P2. Indeed, the PLCC and SROCC values obtained for LIVE-P1 are respectively 0.982 and 0.981, while those obtained for LIVE-P2 are 0.984 and 0.985, respectively. Furthermore, we report the performance of our method according to the size of the training set. Table 5.7 shows the correlations achieved for a training set of size 50%, 70% and 80% using LIVE databases. The partition ratio has a slight impact on the performance. And it does not suffer from an over-fitting problem. The diminution is similar for both datasets. Meanwhile, performance evaluation on Waterloo datasets are not reported in several metrics papers. Table 5.8 shows the state-of-the-art comparison using Waterloo-P1 and Waterloo-P2 databases. In comparison with two FR metrics and four NR metrics including two recently published methods (i.e. Chen [23] and Sun [110]),the proposed approach again outperforms both NR and FR metrics on both Waterloo datasets.

Table 5.7: Performance of the proposed metric using VGG-16 under different train-test partitions on LIVE-P1 and LIVE-P2.

	LIVE-P1			LIVE-P2		
Partition	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
80%-20%	0.981	0.982	3.086	0.984	0.985	1.938
70%-30%	0.980	0.980	3.189	0.982	0.983	2.061
50%-50%	0.976	0.977	3.432	0.977	0.978	2.327

Table 5.8: Overall performance comparison on Waterloo-P1 and Waterloo-P2.

		Waterloo-P1			Waterloo-P2		
Type	Metrics	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
FR	Benoit [12]	0.332	0.332	-	0.165	0.320	-
	Chen [21]	0.457	0.631	-	0.272	0.442	-
	Fezza [38]	0.904	0.898	-	0.890	0.866	-
NR	DECOSINE [127]	0.924	0.943	-	0.914	0.933	-
	Chen [23]	0.923	0.931	5.989	0.922	0.925	7.119
	Sun [110]	-	-	-	0.835	0.840	-
Gray	Proposed	**0.967**	**0.973**	**3.592**	**0.977**	**0.981**	**3.617**

To exhibit the prediction responses against human score (objective scores predicted by our method vs. subjective scores), we show in Fig. 5.4 the scatter plots obtained on the four databases. For all datasets, the distribution of the predicted scores is in accordance with the MOS/DMOS for all the considered degradation types.

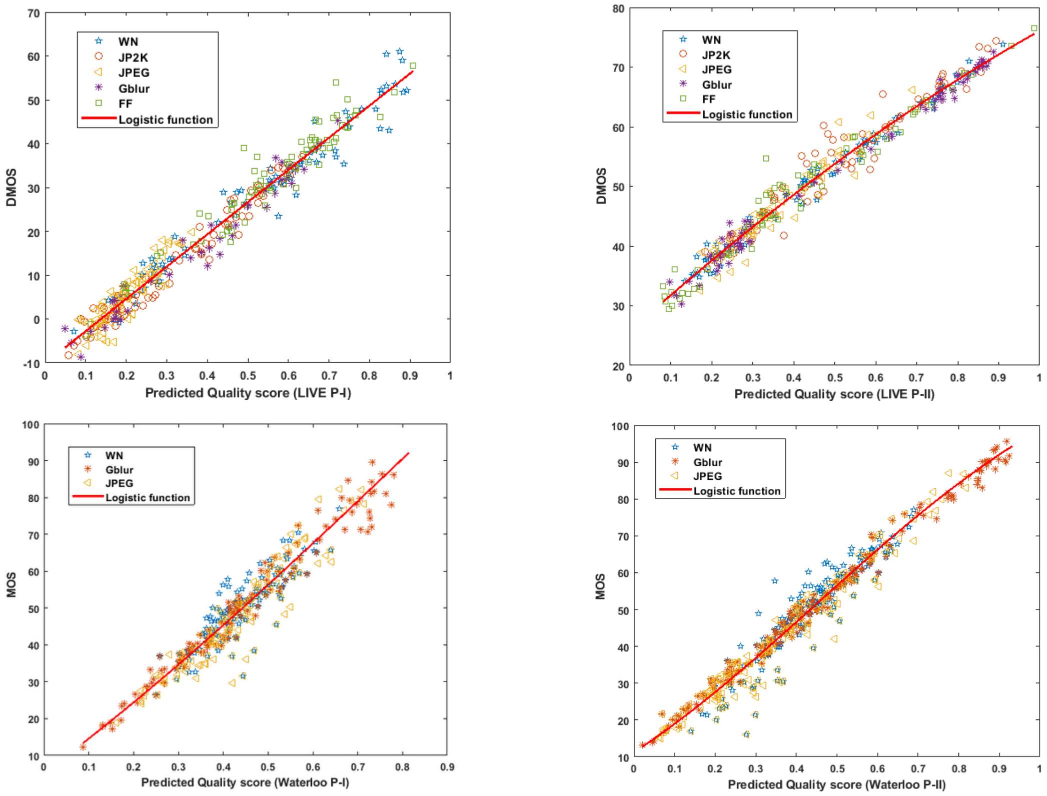


Figure 5.4: Scatter plot of subjective scores against objective scores from the proposed metric on the used four databases.

5.2.9 Performance on individual distortions

The overall performance on the four databases has shown good performance and remarkable consistency. Furthermore, the proposed scheme has been examined on individual distortion types. The performance indexes are computed for each distortion individually. Performance in Tables 5.9, 5.10 and 5.11 indicates that the proposed metric predicts perceptual quality well regardless of types of distortion. Overall, the proposed metric delivers stable performance. On FF subsets, the best accuracy in term of PLCC is achieved by PAD-net metric. In term of SROCC on LIVE-P2, the performance of our metric has achieved the state-of-the-art on all distortion subsets. For Waterloo databases, both the PLCC and SROCC indexes are observed to be above 0.9 on the three distortions JPEG, WN, and BLUR. The highest score has been reached on BLUR distortion. From the used Waterloo and LIVE databases, the metric has reached it highest performance on BLUR. This is also observed in other metrics scores. Usually, the BLUR distortions are easy to detect and they are compared to other forms of distortion such as JPEG one. In the proposed model, the well tuned convolutional layers have given a step further to capture this distortion. On BLUR's distortion over the four datasets, the accuracy of quality assessment was found to be 98% in terms of PLCC.

Table 5.9: PLCC results over five types of Distortions using LIVE-P1 and LIVE-P2.

		LIVE-P1					LIVE-P2				
Type	Metrics	JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN	BLUR	FF
FR	Benoit [12]	0.939	0.640	0.925	0.948	0.747	0.784	0.853	0.926	0.535	0.807
	You [130]	0.877	0.487	0.941	0.919	0.730	0.905	0.830	0.912	0.784	0.915
	Gorley [45]	0.485	0.312	0.796	0.852	0.364	0.372	0.322	0.874	0.934	0.706
	Chen [21]	0.912	0.603	0.942	0.942	0.776	0.834	0.862	0.957	0.963	0.901
	Hewage [50]	0.904	0.530	0.895	0.798	0.669	0.664	0.734	0.891	0.450	0.746
	Bensalma [13]	0.838	0.838	0.914	0.838	0.733	0.666	0.857	0.943	0.907	0.909
RR	RR-BPI [89]	-	-	-	-	-	0.858	0.871	0.891	0.981	0.925
	RR-RDCT [74]	0.918	0.722	0.913	0.925	0.807	0.897	0.748	0.810	0.969	0.910
	Ma [73]	0.940	0.720	0.935	0.936	0.843	0.880	0.765	0.932	0.913	0.906
NR	Akhter [5]	0.905	0.729	0.904	0.617	0.503	0.776	0.786	0.722	0.795	0.674
	Fang [35]	0.911	0.547	0.900	0.903	0.718	0.740	0.764	0.961	0.968	0.867
	DNR-S3DIQE [87]	0.913	0.767	0.910	0.950	0.954	0.865	0.821	0.836	0.934	0.915
	Fezza [38]	-	-	-	-	-	0.936	0.905	0.953	0.974	0.957
	3D-AdaBoost [81]	0.926	0.668	0.941	0.935	0.845	0.835	0.859	0.953	0.978	0.925
	DBN [128]	0.942	0.824	0.954	0.963	0.789	0.886	0.867	0.887	0.988	0.916
	PAD-Net [124]	0.982	**0.919**	0.978	0.985	**0.994**	**0.981**	0.898	0.973	**0.997**	**0.986**
Gray	Proposed	**0.986**	0.906	**0.979**	**0.986**	0.963	0.969	**0.964**	**0.992**	**0.997**	0.982

Table 5.10: SROCC results over five types of distortions using LIVE-P1 and LIVE-P2.

		LIVE-P1					LIVE-P2				
Type	Metrics	JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN	BLUR	FF
FR	Benoit [12]	0.910	0.603	0.930	0.931	0.699	0.751	0.867	0.923	0.455	0.773
	You [130]	0.860	0.439	0.940	0.882	0.588	0.894	0.795	0.909	0.813	0.891
	Gorley [45]	0.015	0.569	0.741	0.750	0.366	0.110	0.027	0.875	0.770	0.601
	Chen [21]	0.888	0.530	0.948	0.925	0.707	0.814	0.843	0.940	0.908	0.884
	Hewage [50]	0.856	0.500	0.940	0.690	0.545	0.598	0.736	0.880	0.028	0.684
	Bensalma [13]	0.817	0.328	0.905	0.915	0.915	0.803	0.846	0.938	0.846	0.846
RR	RR-BPI [89]	-	-	-	-	-	0.776	0.736	0.904	0.871	0.854
	RR-RDCT [74]	0.887	0.616	0.912	0.879	0.696	0.879	0.737	0.732	0.876	0.895
	Ma [73]	0.907	0.660	0.928	0.921	0.792	0.868	0.791	0.954	0.923	0.944
NR	Akhter [5]	0.866	0.675	0.914	0.555	0.640	0.724	0.649	0.714	0.682	0.559
	Zhou [135]	0.856	0.562	0.921	0.897	0.771	0.647	0.737	0.936	0.911	0.798
	Fang [35]	0.880	0.523	0.883	0.523	0.650	0.714	0.709	0.955	0.807	0.872
	DNR-S3DIQE [87]	0.885	0.765	0.921	0.930	0.944	0.853	0.822	0.833	0.889	0.878
	Fezza [38]	-	-	-	-	-	0.927	0.886	0.947	0.928	0.952
	3D-AdaBoost [81]	0.899	0.625	0.941	0.887	0.777	0.842	0.837	0.943	0.913	0.925
	DBN [128]	0.897	0.768	0.929	0.917	0.685	0.859	0.806	0.864	0.834	0.877
	PAD-Net [124]	0.969	0.889	0.968	0.917	**0.996**	0.959	0.882	0.962	0.867	0.945
Gray	Proposed	**0.975**	**0.906**	**0.978**	**0.967**	0.950	**0.963**	**0.957**	**0.988**	**0.983**	**0.972**

Table 5.11: Performance comparison of the proposed metric on individual distortions using Waterloo-P1 and Waterloo-P2 database.

	Waterloo-P1			Waterloo-P2		
Distortion type	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
JPEG	0.951	0.954	4.084	0.968	0.970	4.075
WN	0.915	0.916	3.756	0.940	0.941	4.178
BLUR	0.985	0.987	2.715	0.988	0.995	2.017

Table 5.12 shows the performance of our metric on symmetric and asymmetric distorted stimuli. As can be seen, some metrics totally fail to predict the quality for asymmetric distorted images. They give high correlations for symmetric distorted images (Benoit, You and Bensalma). PAD-Net yields the best performance for symmetric distorted images. The first and the second best correlations for symmetric and asymmetric distorted images have been produced by the proposed approach. According to the Table 5.6; our metric achieves the best global results. Moreover, high accuracy on asymmetric distortions is

more challenging, since most of the existing methods fail.

Table 5.12: SROCC performance for symmetric and asymmetric distorted images on LIVE-P2. Best result of each category is highlighted in bold.

		LIVE-P2		Waterloo-P1		Waterloo-P2	
Method	Type	Symmetric	Asymmetric	Symmetric	Asymmetric	Symmetric	Asymmetric
Benoit [12]	FR	0.860	0.671	-	-	-	-
You [130]		0.914	0.701	0.752	0.571	-	-
Gorley [45]		0.383	0.056	0.566	0.475	-	-
Chen [21]		0.923	0.842	0.924	0.643	-	-
Hewage [50]		0.656	0.496	-	-	-	-
Bensalma [13]		0.841	0.721	-	-	-	-
Akhter [5]	NR	0.420	0.517	-	-	-	-
Fezza [38]		0.928	0.882	0.902	0.869	0.915	0.804
3D-AdaBoost [81]		0.898	0.917	-	-	-	-
PAD-Net [124]		**0.982**	0.954	0.985	**0.978**	-	-
Proposed		0.973	**0.987**	**0.987**	0.967	**0.987**	**0.976**

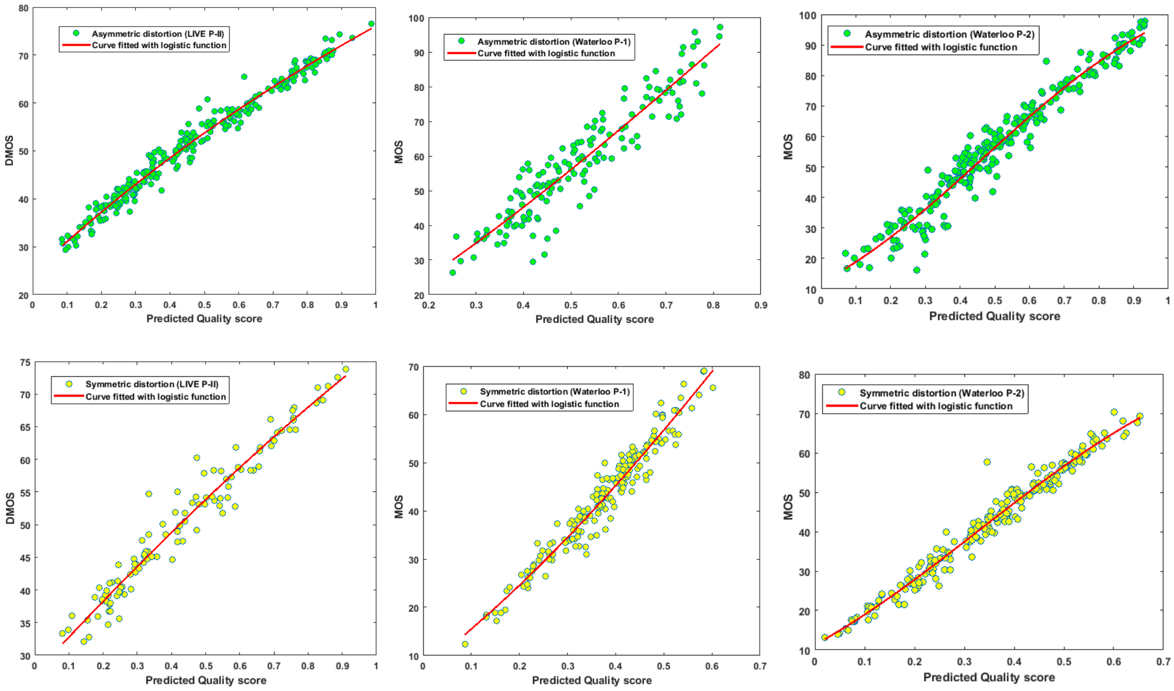


Figure 5.5: Asymmetric and symmetric distortion plots from the four databases using the proposed method.

5.2.10 Cross database performance

Cross-database experiments have been conducted in order to verify the generalization ability of the proposed approach. The implemented tests are shown in Table 5.13. Metrics shown are all NR methods. They have been trained in the former database and tested on the latter.

Table 5.13: PLCC Performance of cross database tests using the four databases. (Expressed as: Train database/Test database.)

Metrics	L-P2/L-P1	L-P1/L-P2	W-P1/W-P2	W-P2/W-P1
DBN [128]	0.869	0.852	-	-
DECOSINE [127]	**0.916**	0.846	**0.842**	**0.873**
3D-AdaBoost [81]	0.892	0.824	-	-
Chen [23]	0.827	0.812	0.806	0.846
Sun [110]	0.899	**0.919**	-	-
PAD-Net [124]	0.915	0.854	-	-
Proposed	0.911	0.851	0.826	0.848

Comparing with the NR metrics, our method has competitive prediction about the quality of stereo pairs despite cross-database tests. DECOSINE, Sun and PAD-net algorithms deliver decent performance in the four cross-database tests, but Sun is the only algorithm which gives performance over 0.9 in term of PLCC in the L1/L2 test. From LIVE datasets, the performance of the other NR algorithms is not as good as the performance of the individual database tests. For instance, Chen and DBN metrics showed good results on the individual database tests where Pearson correlations (PLCCs) of 0.959 and 0.956 have been achieved on LIVE P-1 for Chen and DBN, respectively. They gave low performance scores in the L1/L2 test. PLCC of 0.869 and 0.827 are reported for Chen and DBN respectively. Waterloo datasets have shown lower correlations than LIVE datasets. It is important to notice that which makes Waterloo databases more challenging than LIVE is that they not only include both symmetric and asymmetric distorted pairs like LIVE phase-II. Also, the left and right views of a stereo pair may be distorted by different distortion types. The cross-database tests revealed that the proposed approach ranks third after the two metrics DECOSINE and PAD-net. However on LIVE datasets, the correlation gaps are not profound, 0.003 and 0.005 are the difference values of our metric with PAD-net and DECOSINE respectively.

5.2.11 Influence of distortions on the 3D saliency map

To investigate the impact of the distortions on the computed 3D saliency map from the cyclopean image, we observe the 3D saliency map generated over two different types of distortion, namely JP2K and FF. The cyclopean image is also being spotted on these distortions. Fig. 5.6 displays the computation outputs.

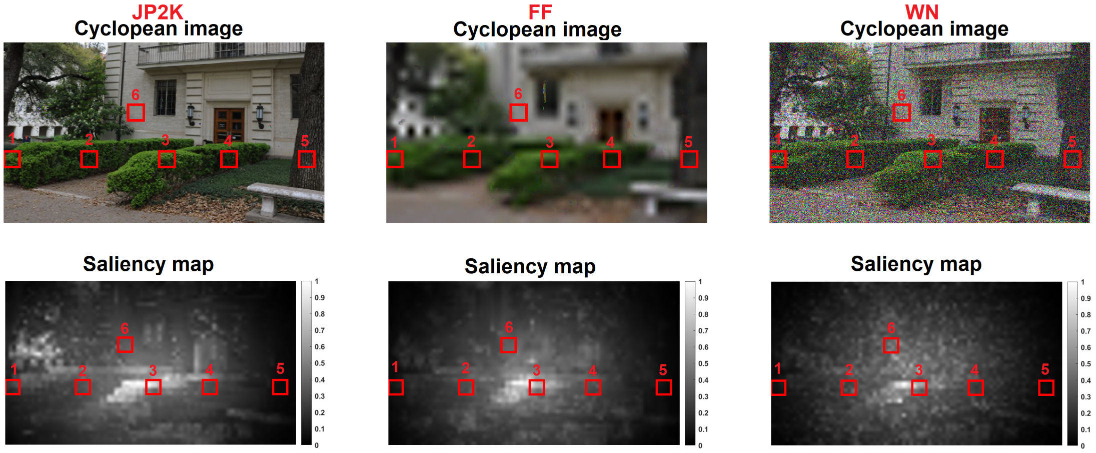


Figure 5.6: Examples of synthesized cyclopean image and 3D saliency map on two different types of distortion.

As can be seen, in each of the synthesized cyclopean image, the quality deformation is clearly stated. It depends on the type of distortion. Meanwhile, the computed 3D saliency maps are very similar despite the variation of distortion. This latter indicates consistency against the degradations that occur in the stereoscopic images. Furthermore, relationship of the saliency value and the error quality prediction are studied. Six patches of the same locations have been selected from each cyclopean and 3D saliency maps as shown in Fig. 5.6. Quality prediction error of each patch P_e and its saliency average P_s computations are as follows:

$$P_e = abs(y - \hat{y}) \tag{5.3}$$

where y is the ground truth quality and $\hat{y}$ is the predicted quality using the proposed metric. The patch saliency average S_a is defined by:

$$S_a = \frac{1}{m.n} \sum_{i=1}^{m} \sum_{j=1}^{n} M(i,j) \tag{5.4}$$

where M is the computed 3D saliency map from previous steps.

Fig. 5.7 shows the obtained curves. On both distortions, it is remarkable to observe the changes of prediction error derived by the saliency. Curves show that the prediction error drops when the saliency patch average increases and vice-versa. In the case of JP2K distortion, patch number three shows that the highest saliency (0.63) is visible at lowest quality prediction error of values (0.004). For FF distortion with the same patch, we note the lowest error (0.068) at the highest saliency value (0.42). Generally, for saliency values above the 0.3 threshold, we find consistency quality prediction errors below 0.15. From these findings, we conclude that the human visual selectivity influences the quality evaluation. This quality evaluation can be improved by saliency information for objective methods.

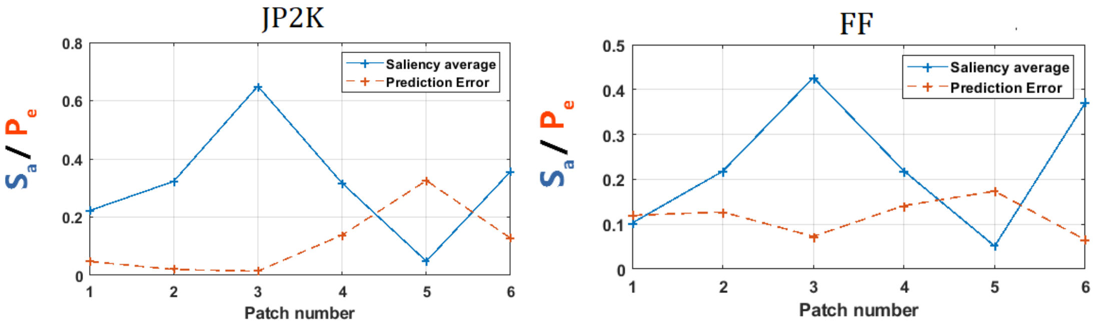


Figure 5.7: Saliency patch average versus quality prediction error for patches from 1 to 6 under JP2K, and FF distortion shown in Fig. 5.6.

5.2.12 Statistical test performance

In order to verify whether our proposed model is statistically better than other metrics. We conducted the T-test against the state-of-the-art metrics with confidence at 90% applied over 10 trials for PLCC and SROCC. This test is one of numerous statistical tests [91] as discussed in chapter 1. The results is statistically superior or worse than the competitive metric in the column, respectively. It is important to remember that the

Table 5.14: T-test results with confidence of 90% of the proposed metric against the others using PLCC, SROCC on LIVE I and II

Database	Index	3D-AdaBoost [81]	Chen [23]	Shen [101]	PAD-Net [124]
LIVE I	PLCC	1	1	1	1
	SROCC	1	1	1	1
LIVE II	PLCC	1	1	1	1
	SROCC	1	1	1	1

value of 1 indicates the superiority of the proposed method, and -1 indicates the opposite. While 0 means that the two metrics are statistically similar. From the tabulated results, we notice that our metric performs statistically better than other NR-SIQA metrics both on LIVE Phase I and II.

5.2.13 Computational complexity

We compare here computational time with the most recent NR-SIQA metrics that incorporate deep learning into their designs. The working platform uses the MATLAB2020a on a computer equipped with Intel(R) Xeon(R) CPU E5-2620 v4 processor at 2.10GHz, 64GB of memory and a NVIDIA Quadro P5000 GPU - 16GB of memory. It should be noted that the other approaches have been tested on various hardware. The test was performed on a stereo image from the LIVE phase II database with a resolution of 640 x 360 pixels.

The run time (in seconds) tests are listed in Table 5.15. It is worth noting that for our model we record the time around 17 seconds for predicting quality score. The results show that PAD-Net [124] only needs around 1 second per image which is significantly lower than other metrics, while metrics in [101, 134] require around 9 and 3 seconds, respectively, to deliver quality ratings. In our approach, the most computationally expensive stage is the cyclopean image construction, since it involves weights computation of the left and right views by performing a multi-scale Gabor filter. Note that the metric in [?] also includes a cyclopean image computation, where this metric records higher run time, around 20 seconds. Therefore, we can observe that metrics which do not require considerable preprocessing, such as cyclopean image computation, are more likely to be faster than others because they mostly use the stereoscopic image directly as input.

Table 5.15: The computation time comparison using NVIDIA Quadro P5000 GPU - 16GB for the proposed method.

Metrics	Shen [101]	StereoQA-Net [134]	PAD-Net [124]	Yang [?]	Proposed
Time (sec.)	8.822	2.377	**0.906**	19.882	16.335

5.2.14 Deep network visualization

In this section, we take a look at what deep convolutional neural network sees from degraded images. We also analyzed the learned convolutional filters and their activation functions. Where we examine which parts of the cyclopean image are most important for our CNN models. To ensure independence output, we have preferred the model trained on LIVE-P2 to observe its behavior on new cyclopean images from LIVE-P1. The test cyclopean images are shown in Fig. 5.6, where only the patch number six is fed to the network. The synthesized cylopean views were formed under different types of distortion: JP2K, WN and FF. The patches are fed to the CNN and then inspect the outputs of activation functions (ReLU) after the first and second convolutional layers. The first two convolution layers produce 64 channels each. Among the 64 channels output from ReLU layer, their mean values are computed and the strongest channel has been selected by indexing the maximum. Fig. 5.8 despite the first and second ReLU layer responses for the input cyclopean patch. As can be seen, where the warmer (closer to 1) regions activate the ReLU function and thus influence the decision of the network. It is remarkable that the first activation function reflects the presence of pixel deformation. The JP2K compression is well known artifact that causes undesirable blocks in the image due to the quantization. This issue is stated in ReLU 1 activation map of JP2K patch that shows the selection of these blocks as a highly important information to pass through the network. As well as for WN and FF cyclopean patches, the ReLU 1 activation function has succeeded to focus on noise and blur artifacts.

While the second activation function (ReLU layer) is controlled by a deeper representation that makes it harder to fully comprehend the outputs. However, for JP2K cyclopean patch, deformed regions cover most of the patch that captures peace of house wall on the scene. For WN, the deformed regions are located around everywhere the wall. From the second ReLU output maps, the warmer regions are somewhat distributed according to the most infected regions in the scene. Meanwhile for FF patch, the spatial information of the wall is less effected since FF is considered as high frequency distortion. Interestingly, the ReLU 2 responses show that the degradation covers the entire wall, which is often the case for FF degradation. Notice that for each patch the activation functions appear diversity as the type of degradation varies. However, the predicted scores for the four patches are similar and follow the human judgments (DMOS).

Overall, we observe the model learns to focus on the pixel deformations to extract a complex quality indicators. Thus, the model can also distinguish between different types of distortion. Based on these findings, we conclude that the deep network extracts high-quality features, which are controlled by the shape and degree of distortion.

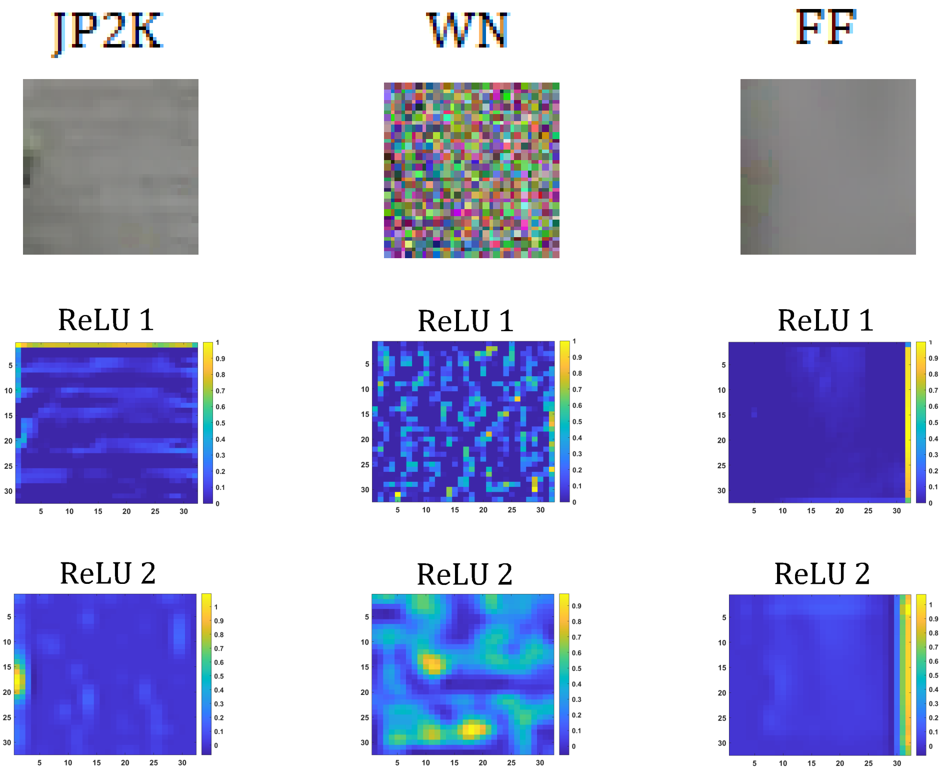


Figure 5.8: The first and second ReLU activation layer outputs (feature map) from a test cyclopean patch for three degradation types.

5.2.15 Conclusion

A new no-reference stereoscopic IQA based on the use of cyclopean image and saliency map has been proposed. The simplicity of the proposed scheme is a benefit for an easy implementation in the multimedia software. Cyclopean image has been introduced to consider asymmetrical distortion, while the saliency aims to focus on the most perceptual relevant regions by selecting relevant patches from the cyclopean image. These patches are then fed as input to a modified version of a pre-trained CNN model to estimate the quality.

We compared five pre-trained models (i.e. AlexNet, VGG16, VGG19, ResNet18 and resent50) and we also show the impact of the saliency selection. The best performance has been obtained with VGG16 for a saliency threshold equals to 0.3. Experimental results have demonstrate the efficiency of the proposed metric since it outperforms all the compared FR and NR SIQA of the state-of-the-art on LIVE and Waterloo databases. Also, the capacity of our method to predict the quality of unknown stereo images has been evaluated.

Chapter 6

General Conclusion and perspectives

6.1 General Conclusion

As demonstrated in the state-of-the-art chapter of this thesis, significant improvements in stereoscopic image quality assessment have been made to date. Human binocular perception research are increasingly being explored for the design of metrics, and promising human simulation algorithms are being implemented and tested. In general, considerable research efforts are being dedicated on enhancing every aspect of stereoscopic content quality. Nevertheless, it is certain that clearly understanding the mechanisms of perceived stereoscopic quality remains a significant difficulty, as does determining the best approach for an accurate quality evaluation that is applicable in a variety of situations.

In this context, the proposed work is an attempt to gain new insights into the properties of human perception and judgment of stereoscopic image quality, as well as to suggest a method for assessing it. Through the chapters of this manuscript, we were therefore able to first analyze the complexity of stereoscopic vision and study the various assessment system types available in the literature. Then, we focused our efforts on better understanding what effects stereoscopic image quality and an observer's judgment criteria. In this concept, we suggest several quality evaluation frameworks. The proposed metrics have been validated after series of tests, where the results show good performance compared to literature. As visual algorithms, we used cyclopean view synthesis and visual saliency to imitate perceptual human characteristics in our quality assessment models. We also employed the most recent deep learning networks/architectures to map perceptual inputs to quality evaluation ratings. This method avoids the time-consuming computation of a stereoscopic quality score.

During our experiments, we discovered that the 2D-IQA metrics are relatively unreliable for stereoscopic images since these approaches do not account for the disparity/depth information of a 3D scene. Another conclusion that could be drawn is that the SIQA metric's primary purpose is HVS simulation, and the better the simulation, the more accurate the metric will be. Finally, we consider our contributions as a step forward in the metrics development for stereoscopic stimulus.

It should be noted that all metrics described in this manuscript were tested on natural scene images, and their effectiveness may vary depending on the type of the content, including different types of capturing, such as two-view or multi-view content.

6.2 Perspectives

The first perspective work that we envisage is a study analysis on the utility of proposed framework for visual discomfort assessment/detection. We believe that the methods suggested in this thesis have room for improvement. On this basis, many different HVS mathematical models might be explored and merged into our concept designs for assessing stereoscopic image quality. Such a mathematical models should be validated by a series of test experiments on different datasets and compared to other existing models. Finally, we believe that our approaches are not restricted to stereoscopic imagery, but also to stereoscopic videos and live 3D streaming broadcasts.

Bibliography

[1] (01/2012), R. B.-. Methodology for the subjective assessment of the quality of television pictures, 2012.

[2] ((04/2008)), T. P. Subjective video quality assessment methods for multimedia applications, 2008.

[3] (10/2019), R. B.-. Methodologies for the subjective assessment of the quality of television images, 2019.

[4] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 11 (2012), 2274–2282.

[5] Akhter, R., Sazzad, Z. P., Horita, Y., and Baltes, J. No-reference stereoscopic image quality assessment. In *IS&T/SPIE Electronic Imaging* (2010), International Society for Optics and Photonics, pp. 75240T–75240T.

[6] Albawi, S., Mohammed, T. A., and Al-Zawi, S. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (2017), Ieee, pp. 1–6.

[7] Appina, B., Khan, S., and Channappayya, S. S. No-reference stereoscopic image quality assessment using natural scene statistics. *Signal Processing: Image Communication 43* (2016), 1 – 14.

[8] Baltes, J., McCann, S., and Anderson, J. Humanoid robots: Abarenbou and daodan. *RoboCup-Humanoid League Team Description* (2006).

[9] Banitalebi-Dehkordi, A., Pourazad, M. T., and Nasiopoulos, P. A study on the relationship between depth map quality and the overall 3d video quality of experience. In *2013 3DTV Vision Beyond Depth (3DTV-CON)* (2013), IEEE, pp. 1–4.

[10] Bartlett, P. L., and Traskin, M. Adaboost is consistent. *Journal of Machine Learning Research 8*, Oct (2007), 2347–2368.

[11] Bay, H., Tuytelaars, T., and Van Gool, L. Surf: Speeded up robust features. In *European conference on computer vision* (2006), Springer, pp. 404–417.

[12] Benoit, A., Le Callet, P., Campisi, P., and Cousseau, R. Quality assessment of stereoscopic images. *EURASIP journal on image and video processing 2008*, 1 (2009), 1–13.

[13] Bensalma, R., and Larabi, M.-C. A perceptual metric for stereoscopic image quality assessment based on the binocular energy. *Multidimensional Systems and Signal Processing 24*, 2 (2013), 281–316.

[14] Blake, R., and Logothetis, N. K. Visual competition. *Nature Reviews Neuroscience 3*, 1 (2002), 13–21.

[15] Blake, R., Westendorf, D. H., and Overton, R. What is suppressed during binocular rivalry? *Perception 9*, 2 (1980), 223–231.

[16] Boev, A., Hollosi, D., and Gotchev, A. Classification of stereoscopic artefacts. *Mobile3DTV Project report, available online at http://mobile3dtv. eu/results* (2008).

[17] Brunnstrom, K., Hands, D., Speranza, F., and Webster, A. Vqeg validation and itu standardization of objective perceptual video quality metrics [standards in a nutshell]. *IEEE Signal processing magazine 26*, 3 (2009), 96–101.

[18] Campisi, P., Le Callet, P., and Marini, E. Stereoscopic images quality assessment. In *2007 15th European Signal Processing Conference* (2007), IEEE, pp. 2110–2114.

[19] Carnec, M., Le Callet, P., and Barba, D. An image quality assessment method based on perception of structural information. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)* (2003), vol. 3, IEEE, pp. III–185.

[20] Chen, L., and Zhao, J. No-reference perceptual quality assessment of stereoscopic images based on binocular visual characteristics. *Signal Processing: Image Communication 76* (2019), 1–10.

[21] Chen, M., Su, C.-C., Kwon, D.-K., Cormack, L. K., and Bovik, A. C. Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Processing: Image Communication 28*, 9 (2013), 1143–1155.

[22] Chen, M.-J., Cormack, L. K., and Bovik, A. C. No-reference quality assessment of natural stereopairs. *IEEE Transactions on Image Processing 22*, 9 (2013), 3379–3391.

[23] Chen, Y., Zhu, K., and Huanlin, L. Blind stereo image quality assessment based on binocular visual characteristics and depth perception. *IEEE Access 8* (2020), 85760–85771.

[24] Chetouani, A. Full reference image quality metric for stereo images based on cyclopean image computation and neural fusion. In *2014 IEEE Visual Communications and Image Processing Conference* (2014), pp. 109–112.

[25] Chetouani, A. Toward a universal stereoscopic image quality metric without reference. In *Advanced Concepts for Intelligent Vision Systems* (Cham, 2015), Springer International Publishing, pp. 604–612.

[26] Corcoran, P., Steinberg, E., Bigioi, P., Drimbarean, A., Zamfir, A., and Florea, C. Image acquisition method and apparatus, May 1 2012. US Patent 8,169,486.

[27] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (2005), vol. 1, Ieee, pp. 886–893.

[28] Daugman, J. G. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision research 20*, 10 (1980), 847–856.

[29] Dendi, S. V. R., Dev, C., Kothari, N., and Channappayya, S. S. Generating image distortion maps using convolutional autoencoders with application to no reference image quality assessment. *IEEE Signal Processing Letters 26*, 1 (2018), 89–93.

[30] Dodgson, N. A. Autostereoscopic 3d displays. *Computer 38*, 8 (2005), 31–36.

[31] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. Support vector regression machines. *Advances in neural information processing systems 9* (1997), 155–161.

[32] Durou, J.-D., and Courteille, F. Integration of a normal field without boundary condition.

[33] Durou, J.-D., Falcone, M., and Sagona, M. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding 109*, 1 (2008), 22–43.

[34] D'Orsi, C. J., Getty, D. J., Pickett, R. M., Sechopoulos, I., Newell, M. S., Gundry, K. R., Bates, S. R., Nishikawa, R. M., Sickles, E. A., Karellas, A., et al. Stereoscopic digital mammography: improved specificity and reduced rate of recall in a prospective clinical trial. *Radiology 266*, 1 (2013), 81–88.

[35] Fang, M., and Zhou, W. Toward an unsupervised blind stereoscopic 3d image quality assessment using joint spatial and frequency representations. *AEU-International Journal of Electronics and Communications 94* (2018), 303–310.

[36] Fang, Y., Chen, Z., Lin, W., and Lin, C.-W. Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Transactions on Image Processing 21*, 9 (2012), 3888–3901.

[37] Feng, X., and Allebach, J. P. Measurement of ringing artifacts in jpeg images. In *Digital Publishing* (2006), vol. 6076, International Society for Optics and Photonics, p. 60760A.

[38] Fezza, S. A., Chetouani, A., and Larabi, M.-C. Using distortion and asymmetry determination for blind stereoscopic image quality assessment strategy. *Journal of Visual Communication and Image Representation 49* (2017), 115–128.

[39] Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A 4*, 12 (1987), 2379–2394.

[40] Fischler, M. A., and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Readings in computer vision*. Elsevier, 1987, pp. 726–740.

[41] Freund, Y., and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences 55*, 1 (1997), 119–139.

[42] Furht, B. A survey of multimedia compression techniques and standards. part i: Jpeg standard. *Real-time imaging 1*, 1 (1995), 49–67.

[43] Gai, S., Da, F., and Dai, X. A novel dual-camera calibration method for 3d optical measurement. *Optics and Lasers in Engineering 104* (2018), 126–134.

[44] Goldstein, E. B., and Brockmole, J. *Sensation and perception.* Cengage Learning, 2016.

[45] Gorley, P., and Holliman, N. Stereoscopic image quality metrics and compression. In *Electronic Imaging 2008* (2008), International Society for Optics and Photonics, pp. 680305–680305.

[46] Hachicha, W., Beghdadi, A., and Cheikh, F. A. Stereo image quality assessment using a binocular just noticeable difference model. In *Image Processing (ICIP), 2013 20th IEEE International Conference on* (2013), IEEE, pp. 113–117.

[47] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015).

[48] Hecht-Nielsen, R. Theory of the backpropagation neural network. In *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[49] Hewage, C., Worrall, S., Dogan, S., Kodikaraarachchi, H., and Kondoz, A. Stereoscopic tv over ip.

[50] Hewage, C., Worrall, S. T., Dogan, S., and Kondoz, A. Prediction of stereoscopic video quality using objective quality models of 2-d video. *Electronics letters 44*, 16 (2008), 963–965.

[51] Hirschmuller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 2, IEEE, pp. 807–814.

[52] Hopf, K. An autostereoscopic display providing comfortable viewing conditions and a high degree of telepresence. *IEEE transactions on circuits and systems for video technology 10*, 3 (2000), 359–365.

[53] IJsselsteijn, W. A., Seuntiëns, P. J., Meesters, L. M., et al. Human factors of 3d displays, 2005.

[54] Ioffe, S., and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[55] Issa, N. P., Trepel, C., and Stryker, M. P. Spatial frequency maps in cat visual cortex. *Journal of Neuroscience 20*, 22 (2000), 8504–8514.

[56] Jiang, Q., Shao, F., Lin, W., and Jiang, G. Learning a referenceless stereopair quality engine with deep nonnegativity constrained sparse autoencoder. *Pattern Recognition 76* (2018), 242–255.

[57] Karimi, M., Soltanian, N., Samavi, S., Najarian, K., Karimi, N., and Soroushmehr, S. R. Blind stereo image quality assessment inspired by brain sensory-motor fusion. *Digital Signal Processing 91* (2019), 91–104.

[58] Kim, J., Ahn, S., Oh, H., and Lee, S. Cnn-based blind quality prediction on stereoscopic images via patch to image feature pooling. In *2019 IEEE International Conference on Image Processing (ICIP)* (2019), IEEE, pp. 1745–1749.

[59] Ko, H., Song, R., and Kuo, C.-C. J. A paraboost stereoscopic image quality assessment (pbsiqa) system. *Journal of Visual Communication and Image Representation 45* (2017), 156–169.

[60] Konolige, K. Small vision systems: Hardware and implementation. In *Robotics research*. Springer, 1998, pp. 203–212.

[61] Korytkowski, M., Rutkowski, L., and Scherer, R. On combining backpropagation with boosting. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on* (2006), IEEE, pp. 1274–1277.

[62] Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. *CoRR abs/1404.5997* (2014).

[63] Lambooij, M., Fortuin, M., Heynderickx, I., and IJsselsteijn, W. Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of imaging science and technology 53*, 3 (2009), 30201–1.

[64] Lambooij, M. T., IJsselsteijn, W. A., and Heynderickx, I. Visual discomfort in stereoscopic displays: a review. In *Stereoscopic Displays and Virtual Reality*

Systems XIV (2007), vol. 6490, International Society for Optics and Photonics, p. 64900I.

[65] Larochelle, H., and Bengio, Y. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning* (2008), pp. 536–543.

[66] Levelt, W. On binocular rivalry (p. 107). *The Hague-Paris: Mouton* (1968).

[67] Liaw, A., Wiener, M., et al. Classification and regression by randomforest. *R news 2*, 3 (2002), 18–22.

[68] Liu, L., Yang, B., and Huang, H. No-reference stereopair quality assessment based on singular value decomposition. *Neurocomputing 275* (2018), 1823–1835.

[69] Liu, Y., Tang, C., Zheng, Z., and Lin, L. No-reference stereoscopic image quality evaluator with segmented monocular features and perceptual binocular features. *Neurocomputing 405* (2020), 126–137.

[70] Lowe, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (1999), vol. 2, Ieee, pp. 1150–1157.

[71] Lv, Y., Yu, M., Jiang, G., Shao, F., Peng, Z., and Chen, F. No-reference stereoscopic image quality assessment using binocular self-similarity and deep neural network. *Signal Processing: Image Communication 47* (2016), 346–357.

[72] Ma, J., An, P., Shen, L., and Li, K. Full-reference quality assessment of stereoscopic images by learning binocular visual properties. *Applied optics 56*, 29 (2017), 8291–8302.

[73] Ma, J., An, P., Shen, L., and Li, K. Reduced-reference stereoscopic image quality assessment using natural scene statistics and structural degradation. *IEEE Access 6* (2017), 2768–2780.

[74] Ma, L., Wang, X., Liu, Q., and Ngan, K. N. Reorganized dct-based image representation for reduced reference stereoscopic image quality assessment. *Neurocomputing 215* (2016), 21–31.

[75] Mehdipour, P., Navidi, I., Parsaeian, M., Mohammadi, Y., MORADI, L. M., REZAEI, D. E., Nourijelyani, K., and Farzadfar, F. Application of gaussian process regression (gpr) in estimating under-five mortality levels and trends in iran 1990-2013, study protocol.

[76] Messai, O., Chetouani, A., Hachouf, F., and Seghir, Z. Deep quality evaluator guided by 3d saliency for stereoscopic images. vol. 2021.

[77] Messai, O., Chetouani, A., Hachouf, F., and Seghir, Z. No-reference stereoscopic image quality predictor using deep features from cyclopean image. vol. 2021.

[78] Messai, O., Hachouf, F., and Seghir, Z. A. Blind stereoscopic image quality assessment using cyclopean view and neural network. In *The fifth IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (2017), IEEE, pp. 196–200.

[79] Messai, O., Hachouf, F., and Seghir, Z. A. Deep learning and cyclopean view for no-reference stereoscopic image quality assessment. In *2018 International Conference on Signal, Image, Vision and their Applications (SIVA)* (2018), IEEE, pp. 1–6.

[80] Messai, O., Hachouf, F., and Seghir, Z. A. Automatic distortion type recognition for stereoscopic images. In *2019 International Conference on Advanced Electrical Engineering (ICAEE)* (2019), pp. 1–4.

[81] Messai, O., Hachouf, F., and Seghir, Z. A. Adaboost neural network and cyclopean view for no-reference stereoscopic image quality assessment. *Signal Processing: Image Communication* (2020), 115772.

[82] Moorthy, A., Su, C.-C., Mittal, A., and Bovik, A. C. Subjective evaluation of stereoscopic image quality. *Signal Processing: Image Communication 28*, 8 (2013), 870–883.

[83] Moorthy, A. K., and Bovik, A. C. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing 20*, 12 (2011), 3350–3364.

[84] Mühlmann, K., Maier, D., Hesser, J., and Männer, R. Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision 47*, 1-3 (2002), 79–88.

[85] Niu, Y., Huang, D., Shi, Y., and Ke, X. Siamese-network-based learning to rank for no-reference 2d and 3d image quality assessment. *IEEE Access 7* (2019), 101583–101595.

[86] of America, M. P. A. 2016 theatrical market statistics report, 2016.

[87] Oh, H., Ahn, S., Kim, J., and Lee, S. Blind deep s3d image quality evaluation via local to global feature aggregation. *IEEE Transactions on Image Processing 26*, 10 (2017), 4923–4936.

[88] Perello Nieto, M. Merging chrominance and luminance in early, medium, and late fusion using convolutional neural networks. G2 pro gradu, diplomityö, 2015-06-10.

[89] Qi, F., Zhao, D., and Gao, W. Reduced reference stereoscopic image quality assessment based on binocular perceptual information. *IEEE Transactions on multimedia 17*, 12 (2015), 2338–2344.

[90] Ramírez-Hernández, L. R., Rodríguez-Quiñonez, J. C., Castro-Toscano, M. J., Hernández-Balbuena, D., Flores-Fuentes, W., Rascón-Carmona, R., Lindner, L., and Sergiyenko, O. Improve three-dimensional point localization accuracy in stereo vision systems using a novel cam-

era calibration method. *International Journal of Advanced Robotic Systems 17*, 1 (2020), 1729881419896717.

[91] Rec, I. P. 1401, methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. *International Telecommunication Union, Geneva, Switzerland* (2012).

[92] Reis Goncalves, N. *Neural computation of depth from binocular disparity.* PhD thesis, University of Cambridge, 2018.

[93] Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., and McGregor, G. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 543*, 2 (2005), 577–584.

[94] Rolland, J. P., and Hua, H. Head-mounted display systems. *Encyclopedia of optical engineering 2* (2005).

[95] Sazzad, Z. P., Yamanaka, S., Kawayokeita, Y., and Horita, Y. Stereoscopic image quality prediction. In *2009 International Workshop on Quality of Multimedia Experience* (2009), IEEE, pp. 180–185.

[96] Schor, C., Wood, I., and Ogawa, J. Binocular sensory fusion is limited by spatial resolution. *Vision research 24*, 7 (1984), 661–665.

[97] Seuntiens, P. Visual experience of 3d tv. *doctor doctoral thesis, Eindhoven University of Technology* (2006).

[98] Seuntiens, P., Meesters, L., and Ijsselsteijn, W. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation. *ACM Transactions on Applied Perception (TAP) 3*, 2 (2006), 95–109.

[99] Shao, F., Li, K., Lin, W., Jiang, G., Yu, M., and Dai, Q. Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties. *IEEE Transactions on Image Processing 24*, 10 (2015), 2971–2983.

[100] Sheikh, H. R., Sabir, M. F., and Bovik, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing 15*, 11 (2006), 3440–3451.

[101] Shen, L., Chen, X., Pan, Z., Fan, K., Li, F., and Lei, J. No-reference stereoscopic image quality assessment based on global and local content characteristics. *Neurocomputing 424* (2021), 132–142.

[102] Si, J., Yang, H., Huang, B., Pan, Z., and Su, H. A full-reference stereoscopic image quality assessment index based on stable aggregation of monocular and binocular visual features. *IET Image Processing* (2021).

[103] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[104] Smith, J. D. The remarkable ibn al-haytham. *The Mathematical Gazette 76*, 475 (1992), 189–198.

[105] Smolic, A., Mueller, K., Stefanoski, N., Ostermann, J., Gotchev, A., Akar, G. B., Triantafyllidis, G., and Koz, A. Coding algorithms for 3dtv—a survey. *IEEE transactions on circuits and systems for video technology 17*, 11 (2007), 1606–1621.

[106] Song, R., Ko, H., and Kuo, C. Mcl-3d: A database for stereoscopic image quality assessment using 2d-image-plus-depth source. *arXiv preprint arXiv:1405.1403* (2014).

[107] Specht, D. F., et al. A general regression neural network. *IEEE transactions on neural networks 2*, 6 (1991), 568–576.

[108] Su, C., Bovik, A. C., and Cormack, L. K. Natural scene statistics of color and range. In *Image Processing (ICIP), 2011 18th IEEE International Conference on* (2011), IEEE, pp. 257–260.

[109] Su, C.-C., Cormack, L. K., and Bovik, A. C. Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation. *IEEE Transactions on image processing 24*, 5 (2015), 1685–1699.

[110] Sun, G., Shi, B., Chen, X., Krylov, A. S., and Ding, Y. Learning local quality-aware structures of salient regions for stereoscopic images via deep neural networks. *IEEE Transactions on Multimedia* (2020).

[111] Suthaharan, S. Support vector machine. In *Machine learning models and algorithms for big data classification*. Springer, 2016, pp. 207–235.

[112] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

[113] Uhr, L. Highly parallel, hierarchical, recognition cone perceptual structures. *Parallel computer vision 4* (1987), 249–292.

[114] Urey, H., Chellappan, K. V., Erden, E., and Surman, P. State of the art in stereoscopic and autostereoscopic displays. *Proceedings of the IEEE 99*, 4 (2011), 540–555.

[115] Wang, J., Da Silva, M. P., Le Callet, P., and Ricordel, V. Computational model of stereoscopic 3d visual saliency. *IEEE Transactions on Image Processing 22*, 6 (2013), 2151–2165.

[116] Wang, J., Rehman, A., Zeng, K., Wang, S., and Wang, Z. Quality prediction of asymmetrically distorted stereoscopic 3d images. *IEEE Transactions on Image Processing 24*, 11 (2015), 3400–3414.

[117] Wang, J., Zeng, K., and Wang, Z. Quality prediction of asymmetrically distorted stereoscopic images from single views. In *2014 IEEE International Conference on Multimedia and Expo (ICME)* (2014), IEEE, pp. 1–6.

[118] Wang, X., Yu, M., Yang, Y., and Jiang, G. Research on subjective stereoscopic image quality assessment. In *Multimedia content access: algorithms and systems III* (2009), vol. 7255, International Society for Optics and Photonics, p. 725509.

[119] Wang, Z., and Bovik, A. C. A universal image quality index. *IEEE signal processing letters 9*, 3 (2002), 81–84.

[120] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13*, 4 (2004), 600–612.

[121] Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (2003), vol. 2, Ieee, pp. 1398–1402.

[122] Westin, C. F. Extracting brain connectivity from diffusion mri [life sciences]. *IEEE Signal Processing Magazine 24*, 6 (2007), 124–152.

[123] William, A. M., and Bailey, D. L. Stereoscopic visualization of scientific and medical content. In *ACM SIGGRAPH 2006 Educators Program* (New York, NY, USA, 2006), SIGGRAPH '06, ACM.

[124] Xu, J., Zhou, W., Chen, Z., Ling, S., and Callet, P. L. Predictive autoencoding network for blind stereoscopic image quality assessment. *arXiv preprint arXiv:1909.01738* (2019).

[125] Yan, J., Fang, Y., Huang, L., Min, X., Yao, Y., and Zhai, G. Blind stereoscopic image quality assessment by deep neural network of multi-level feature fusion. In *2020 IEEE International Conference on Multimedia and Expo (ICME)* (2020), IEEE, pp. 1–6.

[126] Yang, J., Hou, C., Zhou, Y., Zhang, Z., and Guo, J. Objective quality assessment method of stereo images. In *2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video* (2009), IEEE, pp. 1–4.

[127] Yang, J., Sim, K., Gao, X., Lu, W., Meng, Q., and Li, B. A blind stereoscopic image quality evaluator with segmented stacked autoencoders considering the whole visual perception route. *IEEE Transactions on Image Processing 28*, 3 (2018), 1314–1328.

[128] Yang, J., Zhao, Y., Zhu, Y., Xu, H., Lu, W., and Meng, Q. Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network. *Information Sciences 474* (2019), 1–17.

[129] Yegnanarayana, B. *Artificial neural networks.* PHI Learning Pvt. Ltd., 2009.

[130] You, J., Xing, L., Perkis, A., and Wang, X. Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis. In *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA* (2010).

[131] Zhang, W., Qu, C., Ma, L., Guan, J., and Huang, R. Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognition 59* (2016), 176–187.

[132] Zhang, Y., and Chandler, D. M. 3d-mad: A full reference stereoscopic image quality estimator based on binocular lightness and contrast perception. *IEEE Transactions on Image Processing 24*, 11 (2015), 3810–3825.

[133] Zhao, Y., Chen, Z., Zhu, C., Tan, Y.-P., and Yu, L. Binocular just-noticeable-difference model for stereoscopic images. *IEEE Signal Processing Letters 18*, 1 (2011), 19–22.

[134] Zhou, W., Chen, Z., and Li, W. Dual-stream interactive networks for no-reference stereoscopic image quality assessment. *IEEE Transactions on Image Processing 28*, 8 (2019), 3946–3958.

[135] Zhou, W., Qiu, W., and Wu, M.-W. Utilizing dictionary learning and machine learning for blind quality assessment of 3-d images. *IEEE Transactions on Broadcasting 63*, 2 (2017), 404–415.