VISUAL QUALITY ASSESSMENT FOR STEREOSCOPIC VIDEO SEQUENCES

A THESIS SUBMITTED TO
THE GRADUTE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SELİM SEFA SARIKAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL ELECTRONICS ENGINEERING

SEPTEMBER 2011

Approval of the thesis:

VISUAL QUALITY ASSESSMENT
FOR STEREOSCOPIC VIDEO SEQUENCES

submitted by **SELİM SEFA SARIKAN** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen _____
Head of Department, **Electrical Electronics Engineering**

Prof. Dr. Gözde Bozdağı Akar _____
Supervisor, **Electrical Electronics Engineering Dept., METU**

Examining Committee Members:

Prof. Dr. A. Aydın Alatan _____
Electrical Electronics Engineering Dept., METU

Prof. Dr. Gözde Bozdağı Akar _____
Electrical Electronics Engineering Dept., METU

Asst. Prof. Dr. İlkay Ulusoy _____
Electrical Electronics Engineering Dept., METU

Dr. Fatih Kamışlı _____
Electrical Electronics Engineering Dept., METU

Burak Oğuz Özkalaycı, M.Sc. _____
Senior Design Engineer, VESTEK

Date: **14.09.2011**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that as required by these rules and conduct I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Selim Sefa SARIKAN

Signature :

ABSTRACT

VISUAL QUALITY ASSESSMENT FOR STEREOSCOPIC VIDEO SEQUENCES

Sarıkan, Selim Sefa
M.S., Department of Electrical and Electronics Engineering
Supervisor : Prof. Dr. Gözde Bozdağı Akar

September 2011, 69 pages

The aim of this study is to understand the effect of different depth levels on the overall 3D quality and develop an objective video quality metric for stereoscopic video sequences. Proposed method is designed to be used in video coding stages to improve overall 3D video quality. This study includes both objective and subjective evaluation. Test sequences with different coding schemes are used. Computer simulation results show that overall quality has a strong correlation with the quality of the background, where disparity is smaller relative to the foreground. This correlation indicates that background layer is more prone to coding errors. The results also showed that content type is an important factor in determining the visual quality.

Keywords: *3D video, disparity, quality evaluation, segmentation*

ÖZ

ÜÇ BOYUTLU RESİM DİZİLERİ İÇİN GÖRSEL NİTELİK DEĞERLENDİRMESİ

Sarıkan, Selim Sefa
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Gözde Bozdağı Akar

Eylül 2011, 69 sayfa

Bu çalışmanın amacı farklı derinlik seviyelerinin 3B kalitesi üzerindeki nihai etkisini anlamak ve stereoskopik resim dizileri için nesnel bir nitelik değerlendirmesi geliştirmektir. Önerilen yöntem, video kodlama aşamasında kullanılarak toplam 3B video kalitesini yükseltmek için tasarlanmıştır. Bu çalışmada nesnel ve öznel değerlendirmeler birlikte yapılmıştır. Farklı görüntü işleme düzenlerine sahip test dizileri kullanılmıştır. Sonuçlardan bütün kalitenin ön plana göre farklılığın daha az olduğu arka plan ile kuvvetli bağlaşım gösterdiği ortaya çıkmıştır. Bu bağlaşım göstermektedir ki arka plan kodlama hatalarına karşı daha duyarlıdır. Ayrıca sonuçlardan içerik türünün de görsel kaliteyi etkileyen önemli bir etken olduğu görülmüştür.

Anahtar Kelimeler: *3B video, farklılık, kalite değerlendirmesi, parçalara ayırma*

To My Parents

ACKNOWLEDGEMENTS

The author wishes to express his deepest gratitude to his supervisor Prof. Dr. Gözde Bozdağı AKAR for her guidance, advice, criticism, encouragements and insight throughout the research. The author would also like to thank his brother Dr. Alper SARIKAN for his patience, advices, help and technical assistance. The technical assistance of Ramazan Ferhat ÖLGÜN is gratefully acknowledged. Most of all, the author would like to thank to his parents for their invaluable support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVATIONS

DIBR : depth-image-based rendering
FR : full reference
GUI : graphical user interface
HVS : human visual system
MOS : mean opinion score
MSE : mean square error
NR : no reference
PSNR : peak signal-to-noise ratio
RR : reduced reference
SSA : stereo sense assessment
SSIM : structural similarity
VIF : visual information fidelity
VSNR : visual signal-to-noise ratio
VSSIM : video structural similarity

CHAPTER 1

INTRODUCTION

3D technologies are almost becoming an essential part of our daily lives. Many consumer electronic companies are now interested in 3D technologies. Content providers are increasingly streaming 3D video over the Internet. Several broadcasting companies have already launched 3D TV. With this popularity, objective and subjective evaluation of 3D videos are becoming very important.

Even though there exist many 2D quality metrics, it has been known that direct application of these metrics is insufficient for 3D video [1]. In general, there are two approaches for quality evaluation: subjective and objective. Subjective methods utilize human involved testing and give very accurate results but they are time consuming. On the other hand, objective methods do not require human test subjects and they are much faster.

There exist two types of objective quality evaluation. First one is model-based methods. This type of methods tries to model Human Visual System (**HVS**) and estimate distortions. Other methods are feature-based ones. Visual signal-to-noise ratio (**VSNR**) [23] and visual information fidelity (**VIF**) [24] are common examples of model-based image quality metrics. Feature-based methods use signal processing and statistical approaches to estimate the quality. Structural similarity (**SSIM**) [15] and Video Structural Similarity (**VSSIM**) [12] are widely used feature-based image quality indexes.

Model-based approaches require modeling of HVS parts with corresponding signal processing blocks. With the application of proper tuning, model-based methods can give acceptable results. On the contrary, they have the problem of *supra-threshold*. Since the subjective experiments are mostly designed to operate in the *near-threshold* range of vision extrapolation; results in the supra-threshold range is still depend on intuition. On the other hand, feature-based methods use algorithms with weighting values to estimate the effect of distortions on the overall quality. Advantage of feature-based methods is: they are suitable to create effective computer applications. Their disadvantages are need for large data sets for training purposes and subjective testing.

Recently, there are studies on 3D video, both on overall quality assessment and depth perception. Studies in [2] and [3] show the relation between subjective scores and objective metric results using features extracted from 3D videos. There exist other metrics designed specifically for 3D quality measurement. Effects of coding schemes and compression on 3D quality have also been observed. Models in [4] and [7] include the studies made with JPEG coded stereoscopic images. Affects of motion and aspects of disparity are shown in [5] [6]. Studies have been made in [2] [8] [9] [10] and [11] to evaluate 3D quality using information from the depth map.

However, depth and video characteristics are not fully utilized together in all of these proposed metrics. Since HVS depends on binocular depth cues for objects in the near range and highly sensitive to distortions in the structure, all of these information has to be included while proposing a new quality metric for stereoscopic video sequences. The proposed quality evaluation method in this study is designed to include important features mentioned above.

1.1 Scope of the thesis

In this study we propose a new quality evaluation model for stereoscopic videos. The model uses histogram segmentation to divide depth map into visually recognizable planes. Objective scores obtained by the proposed model are compared with the MOS values obtained from subjective tests. Different 3D coding schemes are used to find the effect on the perceived 3D quality.

This study consists of 4 chapters. Chapter 2 presents the method used in quality evaluation used in related studies, depth perception and 3D video types are also discussed in this chapter. In Chapter 3, proposed stereoscopic video quality metric is given and shown that segmentation of depth map and VSSIM can be used to predict 3D video quality. Finally Chapter 4 concludes the study.

CHAPTER 2

BACKGROUND INFORMATION

In this chapter, we will first give information of 3D video representations and then focus on 2D and 3D quality metrics.

2.1 3D Video Representations

Type of 3D video plays an important role in the selection of video processing stages. In this section two main types of the 3D video representations will be investigated. First type is stereo representation which both left and right views are available in the stream and the second one is 2D+depth representation. In this type of video, main stream consists of a single view and depth information. Each of 3D video representation has its own advantages and drawbacks. Stereo content creation requires perfect parallel alignment of the cameras otherwise an unnatural stereo-pair will be presented to observers. This will degrade 3D perception eventually. 2D+depth representation does not require that higher level of adjustment. Also its main stream is 2D compatible. Its disadvantage is that applying a standard 2D compression scheme directly to the depth map could dramatically degrade the 3D effect.

There exist algorithms to switch between two main representations. Depth-image-based rendering (**DIBR**) techniques can be used to find depth map from stereo pair. Reverse is also possible, depth map can be used to generate multiple views with the help of occlusion filling. Two main representations are given in the next figures.

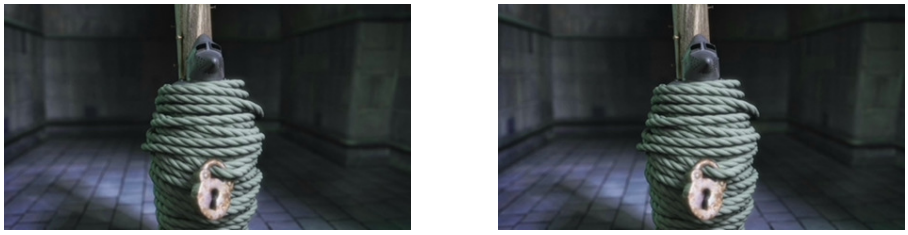

Figure 2.1: Stereo representation

Figure 2.2: 2D+depth representation

2.2 2D Image and Video Quality Metrics

There exist three categories of quality metrics depending on the information required from the reference video. They can be listed as follows:

- full-reference (**FR**)
- no-reference (**NR**)
- reduced-reference (**RR**)

In order to compare each frame of the video under test, FR metrics require the entire reference video. On the other hand NR metrics do not require any information related to reference, distorted video is sufficient for this type of metrics to work on. NR metrics could be more complicated since actual features might not be available due to

distortions. RR metrics stands between these two types of metrics explained, they use some of the features from the reference video but not all of them as in the case of FR metrics.

There has been an increasing need to develop quality measurement techniques that can measure image and video quality automatically. These methods can be used for image and video processing applications such as displaying, analysis, communication, compression, enhancement, watermarking, etc. In general these methods can be used to monitor image and video quality for quality control systems or they can be used to benchmark image and video processing systems and algorithms or they can also be embedded into image and video processing systems to optimize algorithms and parameter settings.

In this section some of the important and widely used image and video quality metrics will be briefly explained.

Peak Signal-to-Noise Ratio (**PSNR**) is the logarithm of the inverse of Mean Square Error (**MSE**) between two image/video frames. MSE is the most commonly used quantitative metric for signal processing performance measurement. It has been known that MSE shows weak performance when used with perceptually important speech signals and images. Although its usage is still being criticized, MSE is still used widely because of its computational efficiency, simplicity and optimization performance.

Main disadvantage of MSE is, its results do not correlate with the results obtained by visual perception. MSE does not include spatial and temporal information of the original signal.

Also it is independent of all the relations between the reference and distorted signals. Image is treated as a whole without any assumption on the error type or location.

Due to lack of these properties two images having the same MSE value, could have different quality perceptions [15], [22]. Calculations of MSE and PSNR are given in (2.1) and (2.2).

$$MSE(X,Y) = \frac{\sum_{j=1}^{N}\sum_{i=1}^{M}\left(X_{i,j} - Y_{i,j}\right)^2}{MN} \tag{2.1}$$

$$PSNR = 20\log\left(\frac{2^n}{MSE}\right) \tag{2.2}$$

'n' stands for the number of bits/pixel in the image.

HVS has the capability of extracting information by seeking to identify and recognize objects. It is highly sensitive for the distortions in the structure and has the ability to compensate for non-structural distortions. In order to reduce the errors caused by MSE by using the important information from the structure, SSIM index is defined [15]. It is composed of three computation stages, luminance comparison, contrast comparison and structural comparison. Detailed description of SSIM calculation will be given in Chapter 3.

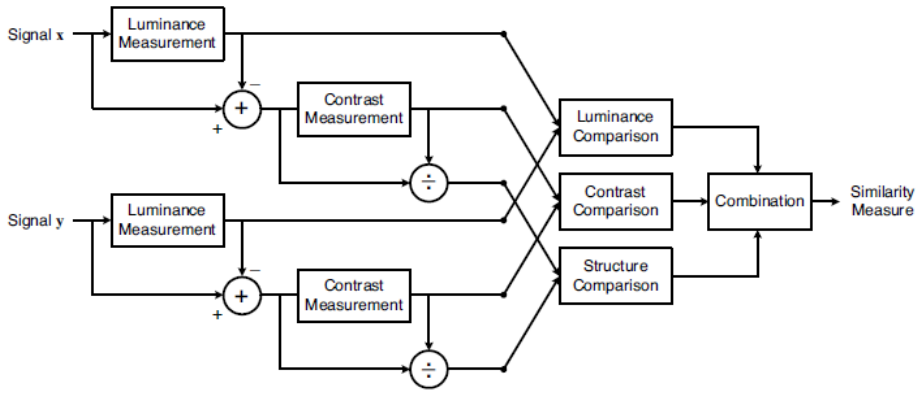

Figure 2.3: SSIM

VSNR [23] is a model-based approach which looks for distortions above the threshold of visual detection. In case of distortions are not visible to human eye, indicating that they are below a certain threshold, metric result will be infinite denoting perfect visual fidelity. Otherwise if distortions are visible, low-level visual property of perceived contrast and the mid-level visual property of global precedence are used. The metric uses multi-scale wavelet decomposition and calculates Euclidean distance in the distortion-contrast space. Its computational complexity and memory requirements are relatively low.

VIF relates signal fidelity to the amount of information that is shared between two signals using an information theoretic approach and modeling HVS and natural image space as well [24]. It also places fundamental limits on the amount of perceptually relevant information that could be extracted from a signal. VIF index exhibits superior performance relative to all other image fidelity measurement algorithms. Its major drawback is algorithmic complexity.

There are strong correlations between adjacent video frames (temporal and spatio-temporal signal structures). Also video contains perceptually important structured motion. Rather than averaging SSIM of individual frames, two adjustments are made to generate a weighted SSIM for video in VSSIM quality index [12]. The first is based on the observation that dark regions usually do not attract fixations, therefore should be assigned smaller weighting values. Second adjustment is by assigning smaller weights for frames with large global motion since image distortions are perceived differently when the background of the video is moving very fast.

2.3 3D Quality Assessment

3D quality metrics is a fairly new research area and there are a few recent papers in the literature. Most of these works start with 2D metrics and try to incorporate information about 3D. Before giving the details of these approaches, we will first give biological aspects of human depth perception since depth perception is one of the key factors determining the 3D video quality.

2.3.1 Depth Perception

Each human eye has its own image of the world. Average distance between the eyes is approximately 6.3 cm [27]. Having two different retinal images creates the binocular disparity [28]. This difference between left and right retinal images carries important information. This information can be of size and depth, such as the relative distance of objects and depth perception of objects within their environment.

In order to obtain depth cues for coding, a wide area having binocular overlapping is used. This process is called binocular vision [29]. Since each eye has its own vantage point separated horizontally, each get different images at the same time. This difference is known as 'binocular stereopsis'. Vision from a single eye can give cues for depth coding, as in the case of perspective, motion parallax, accommodation, relative size, occlusion and relative density. On the other hand, discrimination of different depth levels is enhanced by stereopsis. Existence of two unique retinal images is used by binocular stereopsis to generate depth information. Studies related to binocular stereopsis can be found in [30], [31], [32], [33], [34], and [35].

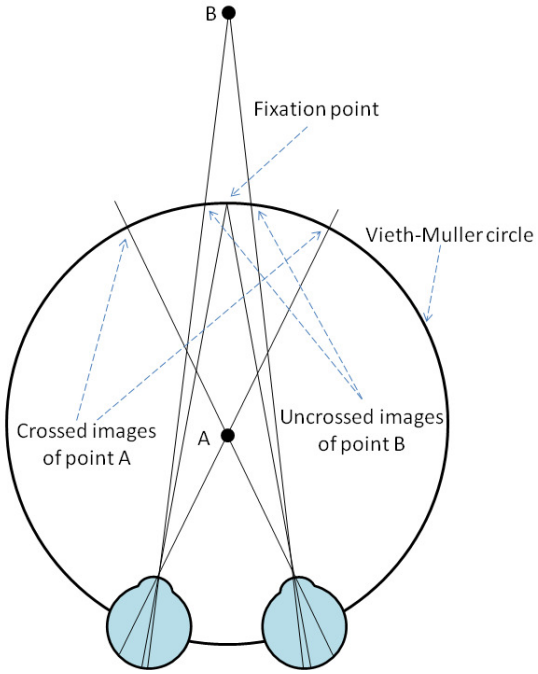

Figure 2.4: Binocular vision

In the presence of available sources of information, humans can estimate distances with high accuracy [36]. Studies made in [37] and [38] states that motion parallax and binocular stereopsis are the most important depth cues. Motion parallax is the result of movements of observer's head and eyes. As the closer and further images are moving relative to each other, the fixation point is automatically adjusted to stay on a specified point. Eyeball movements and head movements are different types of depth cues (motion parallax) used to estimate depth, having similar effects. Depth cues (motion parallax) are affected from the distortions in the temporal domain such as display persistence and motion blur [39]. Having images from the same scene from slightly shifted angles creates binocular depth cues. The process of estimating binocular depth has two stages called vergence and stereopsis. Both eyes adjusted for minimizing the difference between two retinal images is defined as vergence.

Depth cue is extracted from the angle between the two eyes. As the both eyes converge to a fixed point, stereopsis takes place. Stereopsis is the process to estimate depth information from disparity residue. Binocular depth cues have a deep impact on 3D cinema and binocular vision is highly affected from the distortions. Possible results of binocular artifacts can be simulator sickness and nausea. Not all of the people have perfect 3D perception capability, 5% are unable to process binocular depth cues [41], [42], [65], [68].

2.3.2 3D Objective Quality Metrics

In this section major studies on the area of 3D objective quality assessment are given. Metrics given in this section are capable of processing multiview representation and/or 2D+depth.

[2] presents an approach to estimate quality of 3D video by identifying depth planes. They used VQM [45] index for the quality rating of rendered left and right views. This method is mainly concentrated on depth map processing; video is processed as in the case of 2D.

Hewage [3] proposes a RR quality metric for 3D depth map transmission. Binary edge mask is generated using sobel filtering and RR features are extracted from this information. Video content is not used in this study.

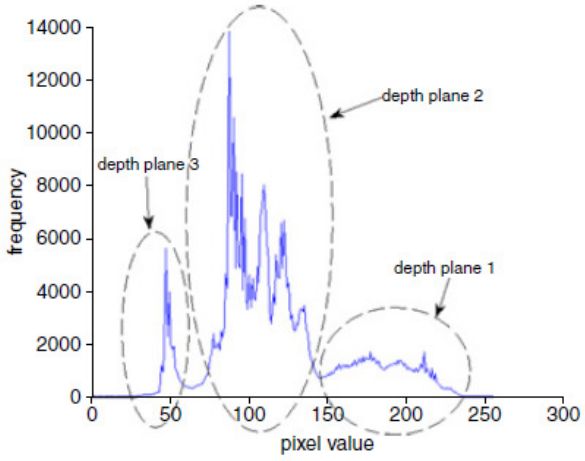

Figure 2.5: Identifying depth planes in [2]

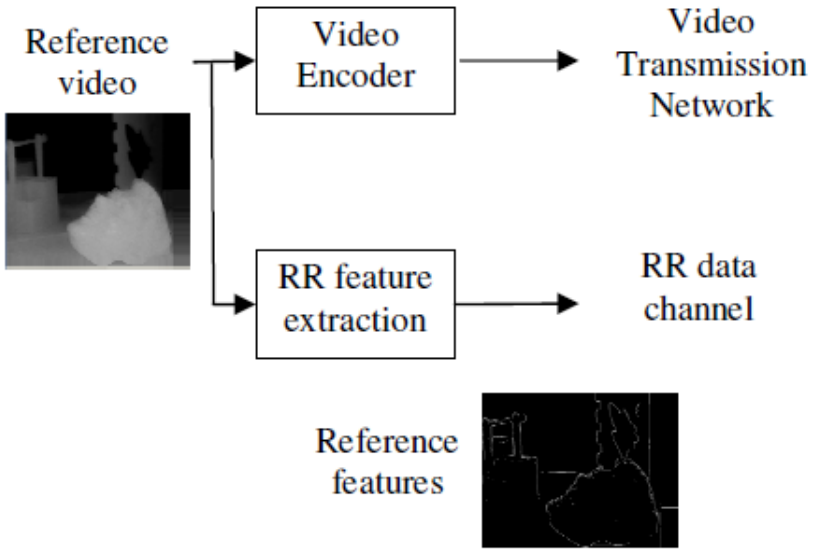

Figure 2.6: RR quality metric in [3]

In [4] a NR quality evaluation model is proposed. This model used features from images as the result of segmentation such as edge, flat and texture regions. This model does not incorporate depth information. Its performance can be seen in Figure 2.7.

In [43], a rapid method which does not use depth map to objectively measure stereo image quality is proposed. Metric has two components: image quality (average of PSNR of two views) and stereo sense assessment (**SSA**) (based on the absolute difference between stereo images). Its performance can be seen in Figure 2.8.

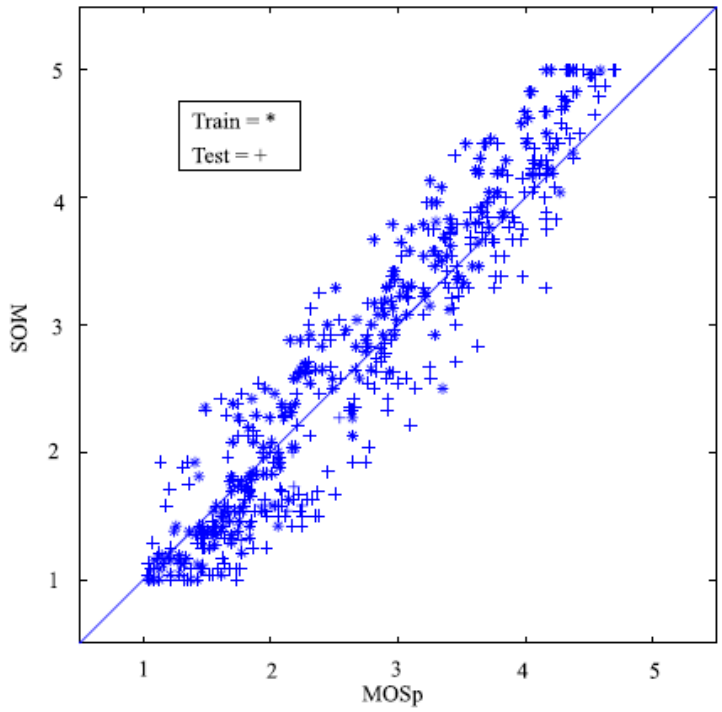

Figure 2.7: Results of NR quality metric in [4]

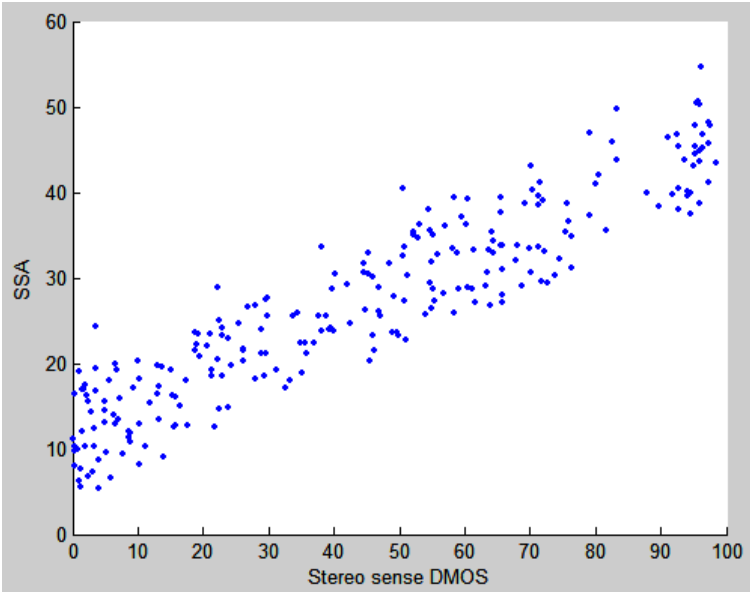

Figure 2.8: Results of FR quality metric in [43]

In [44] a metric is proposed to evaluate DIBR for video plus depth video. It is composed of Color and Sharpness of Edge Distortion (CSED) measure. Color distortion measures the luminance loss of the rendered image compared with the reference, and sharpness of edge distortion calculates a depth-weighted proportion of remaining edge to the original edge (see Figure 2.9).

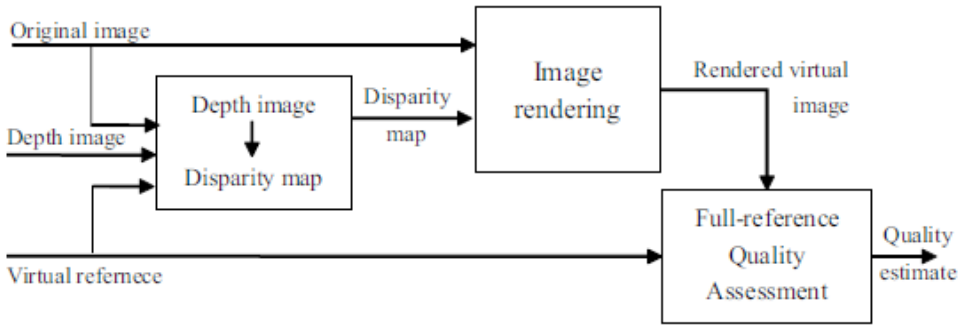

Figure 2.9: Framework used in [44]

2.3.3 3D Subjective Quality Assessment

Subjective assessment of (2D) video quality can be considered to be a mature field. The International Telecommunication Union (ITU) has recommended several methodologies for standard-definition [52], high-definition [53] and low-resolution video [54] [55]. On the other hand, this is not the case for the subjective quality assessment of 3D or stereoscopic video. A first international recommendation was published in 2000 [56]. However, it mostly discusses the way to measure the

stereo acuity of subjects. It also mentions the vergence-accommodation conflict that occurs on most of today's displays due to the fact that flat screens are used. While the accommodation of the HVS focuses on the screen because the objects appear to be most sharp on the display plane, the disparity of the objects between the left and the right eye leads to a convergence of the eyes towards a point in front or behind the display plane. This is an unnatural condition. In the subjective evaluation of 3D images and video sequences, the video quality is closer to the concept of a quality of experience and should be considered to be multi-dimensional: visual quality, depth quality/perception, and comfort. The first dimension may be considered to be the visual quality in the 2D sense because observers usually view a 3D video for the first time in a subjective test, whilst they have a lot of experience with 2D television quality.

The added value of depth was often proposed as a second criterion, and the term naturalness was proposed to express the combination of the perceived depth and the overall quality [57]. Comfort is crucial as it has also been reported that some observers experience visual fatigue with symptoms like eye strain, headache or nausea. This effect is often measured using questionnaires [58]. A recent summary of the causes can be found in [59].

The 3D display itself has a large impact on the stability and reproducibility of the subjective experiment. As the 3D display technology is still advancing, different technologies exist and none can be recommended as a reference. The viewing angle, the field of view, the amount of crosstalk and the brightness are often limiting factors. The International Committee for Display Metrology (ICDM) will soon release a Display Measurement Standard (DMS) to unify the

measurement of display properties [60]. As was mentioned earlier, the 3D content has to be prepared specifically to fit the 3D display, e.g. the depth range has to be adapted. This adjustment depends on the display characteristics and on the viewing distance [61]. Special attention is required on the way the display itself processes the 3D content. Often, crosstalk reduction is applied by the playout program or a format conversion takes place, e.g. from 2D plus depth to nine distinct views, and the rendering artifacts may easily outweigh the added value of depth [62]. ITU-R WP6C is working towards the identification of requirements for the broadcasting and subjective testing of 3DTV [63], whilst ITU-T Study Group 9 added 3D video quality in its scope in 2009 [64]. However, all the issues mentioned previously constitute major challenges in finding a standardized way to characterize and measure the perceived quality of 3D video.

2.3.4 Performance Parameters

Key point to quality metrics is the quality perceived by human observers, MOS. Quality metrics can be characterized by several parameters in terms of its prediction performance, with respect to subjective ratings, presented by [46] as follows: Ability to predict subjective ratings is known as *accuracy*. For data sets with linear relations *Pearson linear correlation coefficient* is used for defining accuracy:

$$r_P = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \tag{2.3}$$

where $\bar{x}$ and $\bar{y}$ denoting means of the respective data sets. In case of nonlinearity is present, a mapping function has to be applied to one of

the data sets. Change in one of the variables affecting the other variable is measured by *monotonicity*.

Monotonicity can be quantified by the *Spearman rank-order correlation coefficient*:

$$r_S = \frac{\sum(\chi_i - \bar{\chi})(\gamma_i - \bar{\gamma})}{\sqrt{\sum(\chi_i - \bar{\chi})^2}\sqrt{\sum(\gamma_i - \bar{\gamma})^2}} \tag{2.4}$$

Where χ_i is the rank of x_i and γ_i is the rank of y_i in the ordered data sets. $\bar{\chi}$ and $\bar{\gamma}$ are their respective midranks.

Consistency is a parameter to measure the number of outliers of all the data points. VQEG proposed consistency to be defined as a data point (x_i, y_i), for which the prediction error is greater than twice the standard deviation [47]:

$$|x_i - y_i| > 2\sigma_{y_i} \tag{2.5}$$

The outlier ratio is then defined as the ratio of the number of outliers and the total number of data points:

$$r_O = N_O / N \tag{2.6}$$

CHAPTER 3

PROPOSED APPROACH

In this chapter detailed explanation of the proposed stereoscopic video quality metric will be given with experimental results.

3.1 Overview

Many of the objective quality metrics designed for 3D content are not fully capable of processing both depth perception and 2D quality together. They have limitations and deficiencies in the generation of meaningful results comparable with MOS and coverage of motion properties for video. In this study a new quality evaluation model for stereoscopic videos is proposed to outcome these limitations. Proposed model can be used with both stereo and 2D+depth represented videos. The methods discussed in [2] and [3] are possible solutions of the system, but these approaches are limited to image content only. Some methods are specifically designed for a single type of 3D video, only for stereo pair or 2D+depth. With DIBR techniques both 3D video types can be processed either.

The proposed method takes two inputs: original video and distorted video, thus it is an FR metric. In case of depth map absence, depth map will be rendered from stereo video input. Next step will be generation and segmentation of depth map histogram. Finally VSSIM index will be combined with the features extracted from the histogram to take the effects of depth perception into account.

By processing statistical information of the image content proposed metric is feature- based. As a consequence of designing a feature-based quality metric, in order to comply with the MOS values; weights are tuned. Calibration of weight values are explained in section 3.4. A computer application is designed for testing the proposed objective quality metric with the ability to change internal parameters. Snapshots and usage of the test application is given in section 0. During computer simulations 4 set of test sequences with different properties are used. Test sequences used in computer simulations, graphs and results obtained by comparison with the subjective test scores are also given in this chapter.

The block diagram of the proposed model is given in Figure 3.1.

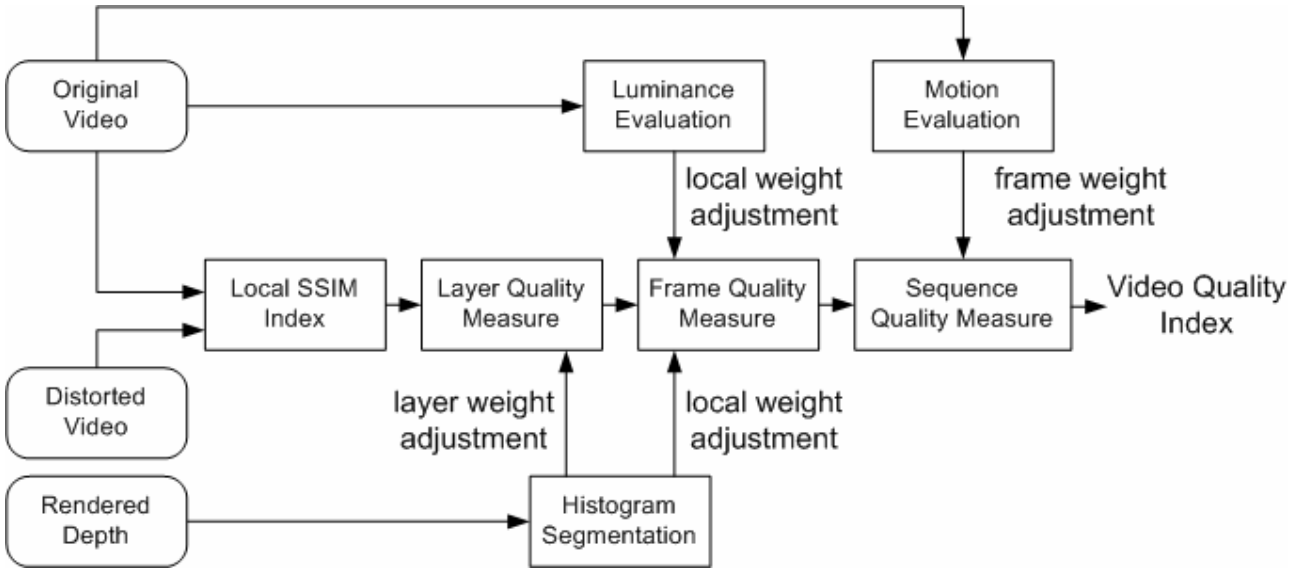

Figure 3.1: Block diagram of the proposed method

3.2 Depth Processing

Depth processing is needed only if the input video is represented as stereo. It consists of 2 stages: Initially, rendering is completed by using a stereo correspondence algorithm. At the second stage, depth map histogram is generated and segmented using a fine to coarse segmentation procedure. Each stage is given in the following subsections.

3.2.1 Rendering

Graph cuts stereo correspondence algorithm (KZ1) [13] is used for estimating depth. Graph cut algorithm as explained in [66] is an optimization solution for flow networks. Its usage in stereo is finding pixels corresponding to the same 3D feature and recovering depth information by means of triangulation. Difficulty is to find matching pixels with high accuracy. Using an energy minimization framework this difficulty can be solved without trapping in local minima. Image shown in Figure 3.4 is generated using the left and right images.

Figure 3.2: Original Left Image

Figure 3.3: Original Right Image

Figure 3.4: Generated depth map

3.2.2 Histogram Segmentation

Histogram of the calculated depth map is plotted and divided into non-overlapping layers using the non parametric approach described in [14]. Segmentation algorithm follows the fine-to-coarse direction. It starts with all the local minima points and rejects inappropriate points to avoid under-segmentation and over-segmentation.

The motivation behind the usage of this method for histogram segmentation is, it does not require the estimate of final number of modes. This property is very important since number of layers can vary by changing video content.

Algorithm is as follows:

Fine to Coarse Segmentation Algorithm

The following algorithm segments a 1D-histogram without a priori assumptions about the underlying density function [49].

h : a discrete histogram

N : number of samples

L : number of bins

$h(i)$: value of h in the bin i

For each interval $[a,b]$, $h(a,b) = h(a) + h(a+1) + \ldots + h(b)$

step 1

Let $S = \{s(0), \ldots, s(n)\}$ be the list of all the local minima, plus the endpoints 1 and L of the histogram h.

n is termed length of the segmentation.

step 2

Choose i randomly in $[1 : length(S) - 1]$

For each t in the interval $[s(i-1) : s(i+1)]$, compute the increasing Grenander [67] estimator h_c of h on the interval $[s(i-1) : t]$ and the decreasing Grenander estimator h_d of h on the interval $[t : s(i+1)]$; Pooling Adjacent Violators algorithm can be used for calculation.

Let $L_c = t - s(i-1) + 1$ be the length of the interval $[s(i-1):t]$ and $N_c = h(s(i-1),t)$ its number of samples (respectively $L_d = s(i+1) - t + 1$ and $N_d = h(t,s(i+1))$ the length and number of samples of $[t:s(i+1)]$).

For each sub-interval $[a,b]$ of $[s(i-1):t]$, compute $NFA_c([a,b])$

$$NFA_c([a,b]) = \frac{L_c(L_c+1)}{2} \mathrm{B}(N_c, h(a,b), \frac{h_c(a,b)}{N_c}) \quad (3.1)$$
$$if\ h(a,b) \geq h_c(a,b)$$

$$NFA_c([a,b]) = \frac{L_c(L_c+1)}{2} \mathrm{B}(N_c, N_c - h(a,b), 1 - \frac{h_c(a,b)}{N_c}) \quad (3.2)$$
$$if\ h(a,b) < h_c(a,b)$$

where $\mathrm{B}(n,k,p)$ denotes the binomial tail

$$\mathrm{B}(n,k,p) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} \quad (3.3)$$

If $NFA_c([a,b]) \leq 1/2$, then $[a,b]$ is said to be a meaningful rejection for the increasing hypothesis on $[s(i-1):t]$.

In the same way, for each sub-interval $[a,b]$ of $[t:s(i+1)]$, compute $NFA_d([a,b])$.

$$NFA_d([a,b]) = \frac{L_d(L_d+1)}{2} \mathrm{B}(N_d, h(a,b), \frac{h_d(a,b)}{N_d}) \quad (3.4)$$
$$if\ h(a,b) \geq h_d(a,b)$$

$$NFA_d([a,b]) = \frac{L_d(L_d+1)}{2} \mathrm{B}(N_d, N_d - h(a,b), 1 - \frac{h_d(a,b)}{N_d}) \qquad (3.5)$$

$$if\ h(a,b) < h_d(a,b)$$

If $NFA_d([a,b]) \le 1/2$ then $[a,b]$ is a meaningful rejection for the decreasing hypothesis on $[t : s(i+1)]$.

If there exists t in $[s(i-1) : s(i+1)]$ with no rejection of the increasing hypothesis on $[s(i-1) : t]$ and no rejection of the decreasing hypothesis on $[t : s(i+1)]$, then $[s(i-1) : s(i+1)]$ is said to follow the unimodal hypothesis. In this case, merge the intervals $[s(i-1) : s(i)]$ and $[s(i) : s(i+1)]$ and remove $s(i)$ from S.

step 3

Repeat *step 2* until no more pair of successive intervals follows the unimodal hypothesis.

step 4

Repeat *step 2* and *step 3* with the unions of j segments, j going from 3 to $length(S)$.

When N is too large, the computation of the binomial tail can be replaced by a large deviation approximation

$$-\frac{1}{N}\log \mathrm{B}(n,k,p) \sim \frac{k}{N}\log\frac{k}{Np} + (1-\frac{k}{N})\log\frac{1-k/N}{1-p} \qquad (3.6)$$

$$if\ \frac{k}{N} > p\ with\ k\ fixed\ and\ N \to \infty$$

The Pooling Adjacent Violators algorithm

The following algorithm [50] [51] computes the decreasing Grenander estimator of a histogram r (apply it to $-r$ to compute the increasing estimator). In order to compute the Grenander decreasing estimator of a histogram h on a given interval l, replace r in the following by the histogram h restricted to l.

Repeat the following operation until you get a decreasing distribution:

For each interval $[i, j]$ on which r is increasing, i.e. $r(i) \le r(i+1) \le \ldots \le r(j)$ and $r(i-1) \ge r(i)$ and $r(j+1) < r(j)$, replace the values $\{r(i), \ldots, r(j)\}$ in r by the mean value on the interval: $\frac{r(i) + \ldots + r(j)}{(j-i+1)}$.

Histogram shown in Figure 3.5 is taken from Butterfly test sequence. Segmented sections are partitioned by red and blue lines respectively. Horizontal green line is showing the average. Numbers on the upper left corner are running values of mean and standard deviation.

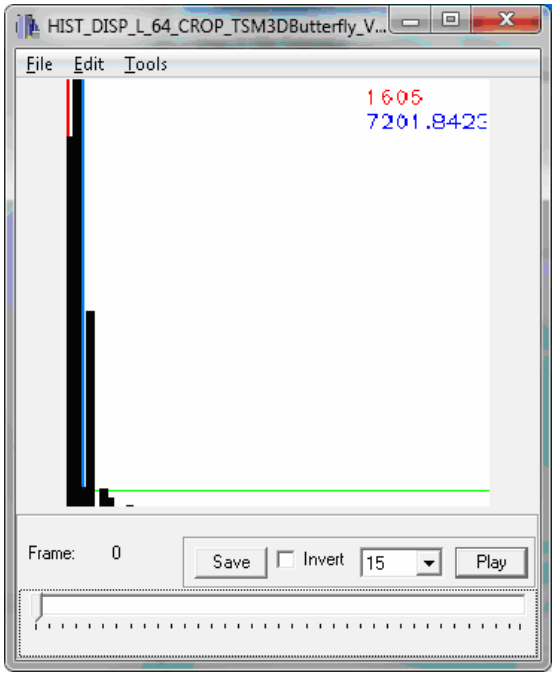

Figure 3.5: Depth map histogram

Extracted layers are shown in Figure 3.7. Corresponding partitioning line colors are used from histogram segmentation. Background is marked with red color and foreground is painted with blue. Shape of the rabbit can easily be selected from the image. Start value of histogram which is initially '0' is stored as z_{far} level. Value of the last bin in the histogram is stored in z_{near} value. These values will be used for optimization purposes in the following sections.

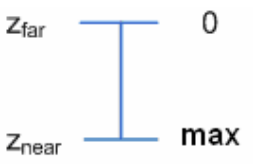

Figure 3.6: Depth level convention

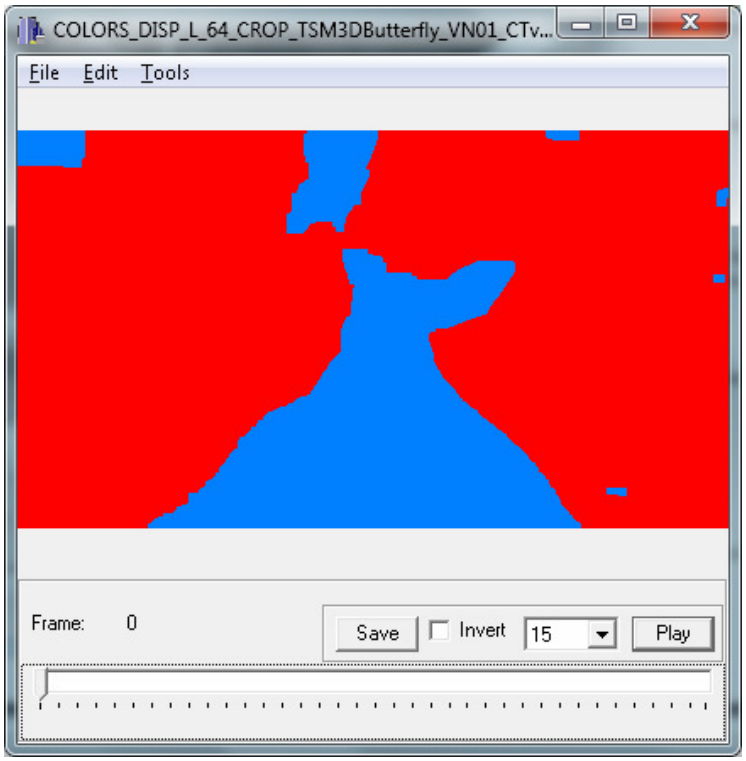

Figure 3.7: Colored layers

Background and foreground layers are extracted with the masks generated from segmented histogram. Corresponding layers can be seen in Figure 3.8 and Figure 3.9. Without using the selected histogram segmentation algorithm, unwanted layers will be generated which will reduce the accuracy of the results.

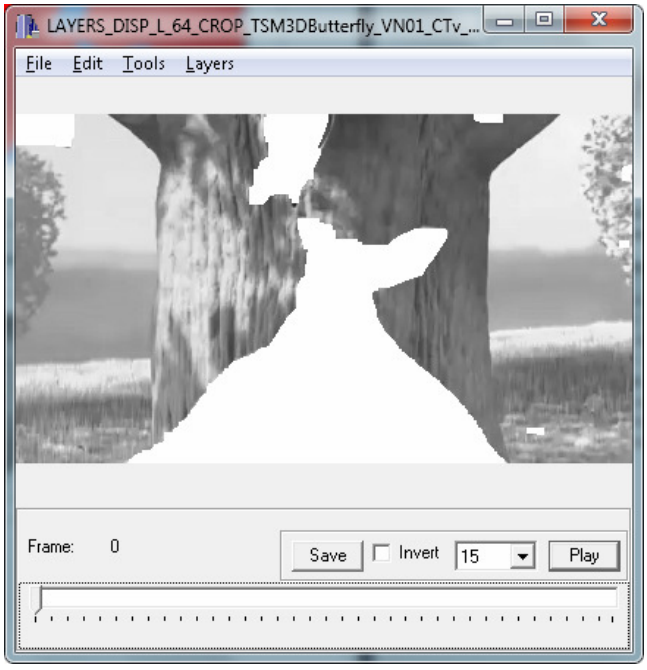

Figure 3.8: Background layer

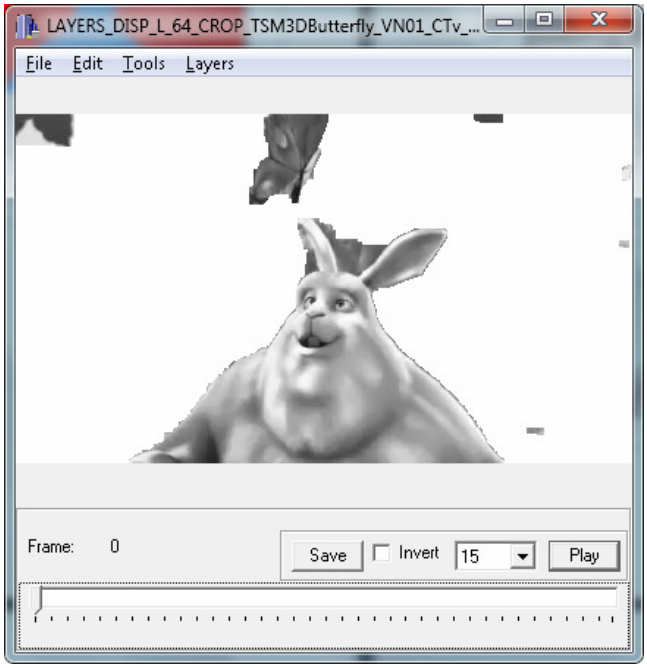

Figure 3.9: Foreground layer

3.3 Quality Estimation

The proposed metric is based on VSSIM [12] with three major calculation steps consisting of layer level, frame level and sequence level. Before detailed explanation of each level, SSIM and VSSIM calculation steps are briefly explained in the following subsections.

3.3.1 SSIM Index Calculation

Given **x** and **y** are two image signals. Three comparisons are performed to measure similarity between these two signals.

- Luminance : $l(x, y)$
- Contrast : $c(x, y)$
- Structure : $s(x, y)$

For luminance comparison, mean intensity is used.
It is represented by μ_x.

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.7}$$

For signal contrast estimation, standard deviation (σ_x) is used, which is given by:

$$\sigma_x = \sqrt{\left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)} \tag{3.8}$$

Normalized signals having unit standard deviation are included in structure comparison.

$$(x-\mu_x)\big/\sigma_x \,,\ (y-\mu_y)\big/\sigma_y \tag{3.9}$$

Overall similarity score is the combination of theses three components.

$$S(x,y) = f\left(l(x,y), c(x,y), s(x,y)\right) \tag{3.10}$$

All of the components used in the above equation are relatively independent.
First stage is luminance comparison, defined by:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C1} \tag{3.11}$$

C_1 is a constant used for biasing, in order to avoid division be zero errors. Which is given by:

$$C_1 = (K_1 L)^2 \tag{3.12}$$

where L, is the maximum pixel value and K_1 is a constant ($K_1 << 1$).
(3.11) can be rearranged as in (3.13)

$$l(x,y) = \frac{2\dfrac{\mu_y}{\mu_x}}{1+\left(\dfrac{\mu_y}{\mu_x}\right)^2 + \dfrac{C_1}{\mu_x^2}} \tag{3.13}$$

Contrasts are compared as follows:

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2} \tag{3.14}$$

As in the case of C_1, C_2 is a constant used for biasing, in order to avoid division be zero errors ($K_2 << 1$).

$$C_2=(K_2L)^2 \tag{3.15}$$

Structure comparison is defined as follows:

$$s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \tag{3.16}$$

σ_{xy} is the cross correlation between two image signals.

$$\sigma_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\mu_x\right)\left(y_i-\mu_y\right) \tag{3.17}$$

(3.11), (3.14) and (3.16) are combined together to obtain final SSIM index.

$$\begin{aligned} &SSIM(x,y)=\left[l(x,y)\right]^{\alpha}.\left[c(x,y)\right]^{\beta}.\left[s(x,y)\right]^{\gamma} \\ &\alpha>0,\beta>0,\gamma>0 \end{aligned} \tag{3.18}$$

For a simpler expression parameters are adjusted as in (3.19).

$$\alpha = \beta = \gamma = 1 \quad , \qquad C_3 = C_2 / 2$$
$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.19}$$

Initially SSIM [15] index is calculated for each layer. Sampling window size is selected as 8x8. Sampling windows located on the layer boundaries are masked so that interference is removed between adjacent layers. Only Y component is used in order to reduce computational requirements.

3.3.2 VSSIM Index Calculation

VSSIM diagram is shown in Figure 3.10. It uses sampling window variable in order to reduce computational requirements. Sampling window count per video frame is represented by R_S. Also SSIM index is calculated for each color channel independently (3.20).

$$SSIM = W_Y\, SSIM^{Y} + W_{Cb}\, SSIM^{Cb} + W_{Cr}\, SSIM^{Cr}$$
$$W_Y = 0.8, W_{Cb} = 0.1, W_{Cr} = 0.1 \tag{3.20}$$

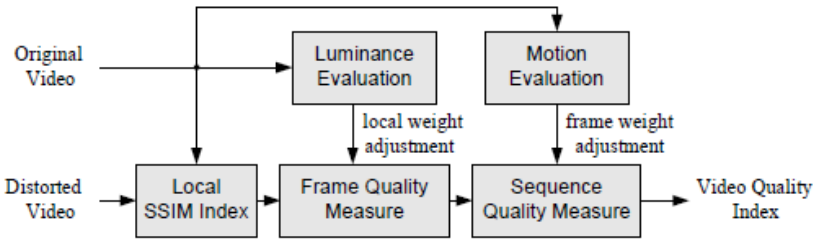

Figure 3.10: VSSIM

Local quality index values are combined in (3.21).

Q_i is quality index measure of i^{th} frame

w_{ij} is the weighting value given to the j^{th} sampling window of the i^{th} frame

$$Q_i = \frac{\sum_{j=1}^{R_S} w_{ij} SSIM_{ij}}{\sum_{j=1}^{R_S} w_{ij}} \tag{3.21}$$

Overall quality is found as

$$Q = \frac{\sum_{i=1}^{F} W_i Q_i}{\sum_{i=1}^{F} W_i} \tag{3.22}$$

First weighting value is applied to dark regions which are not attractive to human eye.

$$w_{ij} = \begin{cases} 0 & \mu_x \le 40 \\ (\mu_x - 40)/10 & 40 < \mu_x \le 50 \\ 1 & \mu_x > 50 \end{cases} \tag{3.23}$$

Second weighting is applied to include effects of global motion. Blurring is more distinctive in still rather than fast moving background. Block based motion is estimated for each sampling window.

Resultant weight is given as

$$M_i = \frac{\left(\sum_{j=1}^{R_s} m_{ij}\right) / R_S}{K_M} \qquad (3.24)$$
$$K_M = 16$$

Frame weights are adjusted by

$$W_i = \begin{cases} \sum_{j=1}^{R_S} w_{ij} & M_i \le 0.8 \\ ((1.2 - M_i)/0.4) \sum_{j=1}^{R_S} w_{ij} & 0.8 < M_i \le 1.2 \\ 0 & M_i > 1.2 \end{cases} \qquad (3.25)$$

3.3.3 Layer Quality Estimation

At the layer level of quality assessment local SSIM values are combined with a weighted summation. Let $SSIM_{ijk}$ denote the SSIM index value of the k^{th} sampling window of the j^{th} layer of the i^{th} frame.

Main aim of this step in the calculation is to find the effect of individual layers on the overall quality; it includes both 2D image quality and provides a basis for depth perception. Selection of weight values are explained in the proceeding subsections. The layer quality index is given as:

$$L_{ij} = \frac{\sum_{k=0}^{n_{ij}} w_{ijk} SSIM_{ijk}}{\sum_{k=0}^{n_{ij}} w_{ijk}} \qquad (3.26)$$

L_{ij} denotes the quality index measure of the j^{th} layer of the i^{th} frame, and w_{ijk} is the weighting value of the k^{th} sampling window of the j^{th} layer of the i^{th} frame. n_{ij} is the number of sampling windows in the corresponding layer.

3.3.4 Frame Quality Estimation

In order to merge results from different layers, each layer has to be weighted accordingly. At the second level of calculation, layer quality measures are combined into a frame level quality measure.

Main aim of this step is for only depth perception. Each layer weight is adjusted to be compatible with the MOS scores.

$$F_i = \frac{\sum_{j=0}^{N_i} m_{ij} L_{ij}}{\sum_{j=0}^{N_i} m_{ij}} \tag{3.27}$$

F_i denotes the quality index measure of the i^{th} frame of the 3D video, m_{ij} is the weight applied to the j^{th} layer of the i^{th} frame. N_i is the number of layers in the corresponding frame.

3.3.5 Video Quality Estimation

Videos have motion characteristics in addition to 2D images. This effect should also be compensated while estimating quality. At the final step frame level quality index measures are summed to a single video quality index measure as follows:

$$Q = \frac{\sum_{i=0}^{I} W_i F_i}{\sum_{i=0}^{I} W_i} \tag{3.28}$$

W_i denotes the weighting value of the i^{th} frame in the sequence and I is total number of frames.

3.4 Calibration of Weight Values

An important stage for an objective quality metric is how well its results are correlated with the subjective tests. To increase this correlation function parameters are adjusted accordingly.

The proposed model includes two weighting factors. First weight is applied at the layer level so that sampling windows at the layer boundaries have larger weighting values. Depth value z is used for this purpose and weight is applied as follows:

$$w_{ijk} = \begin{cases} 1 + \dfrac{z_{near} - z_{ij}}{z_{near} - z_{far}} & boundary \\ 1 & otherwise \end{cases} \tag{3.29}$$

Second weighting includes affect of depth level. As the distance is increased from the z_{near} level, overall quality is more affected. With this weighting value applied background is assigned to larger weight values. S_{ij} is the size of the j^{th} layer in the i^{th} frame, given in pixel count.

$$m_{ij} = S_{ij} \frac{z_{near} - z_{ij}}{z_{near} - z_{far}} \tag{3.30}$$

At the last step, motion information used in VSSIM quality index is used. Motion vector length of the frame is found exhaustive block matching algorithm. W_i is given as follows:

$$M_i = \frac{\sum_{k=0}^{n_i} m_{ik}}{K_M} \tag{3.31}$$

$$W_i = \begin{cases} \sum_{k=0}^{n_i} m_{ik} & M_i \le 0.8 \\ \dfrac{((1.2 - M_i)/4)}{\sum_{k=0}^{n_i} m_{ik}} & 0.8 \le M_i \le 1.2 \\ 0 & M_i \ge 1.2 \end{cases} \tag{3.32}$$

m_{ik} is the motion vector length of the k^{th} sampling window of the i^{th} frame and n_i is the number of sampling windows in the i^{th} frame. K_M is the normalization constant from VSSIM implementation In order to calculate global motion, W_i is calculated at the frame level rather than the layer level.

3.5 Computer Simulations

Main aim of the experiments is to use disparity information to find the effect of each depth layer on the overall quality. Experiments are performed using 3D videos with different coding schemes and levels to measure the performance *of* the proposed full-reference 3D video quality model.

Subjective tests are performed with 87 non professional participants (age: 16-37 years, mean: 24). All subjects were screened to normal visual acuity. Evaluation is performed using Absolute Category Rating (ACR) according to ITU-R P.910. 11-range unlabeled scale is used for overall quality evaluation with randomized stimuli order. An auto stereoscopic (NEC) 3.5" display with 428x240 resolution is used for display purposes. [21]

Properties and bit rates of the test videos are given in Table 1 and Table 2. Correlation between subjective and objective scores is found using R-Squares method which shows results are reliable if greater than 0.7. R-Squares quality ratings are shown in Table 3.

Table 1: Properties of Test Sequences

Sequence	Movement		Complexity		Size
	Camera	Object	Structural	Depth	
Horse	none	low	high	medium	428x240
Car	high	low	medium	high	428x240
Bullinger	none	low	low	low	428x240
Butterfly	none	high	high	medium	428x240

Table 2: Bitrates of the Test Sequences

Quality Level	Butterfly	Car	Horse	Bullinger
Low (kbit/sec)	143	130	160	74
High (kbit/sec)	318	378	450	160

Table 3: Quality Ratings for Test Sequences

Sequence	R-square			
	Left		Right	
	Back.	Fore.	Back.	Fore.
Butterfly	0.7695	0.7706	0.9817	0.9598
Horse	0.7138	0.7436	0.9390	0.9207
Car	0.9210	0.9128	0.9806	0.8275
Bullinger	0.8891	0.9214	0.3135	0.6624

3 different 3D video coding schemes with 2 different levels are tested. Mixed Resolution Stereo Coding (MRSC) coding each view with a different resolution where quality is dominated by the lower quality view Multi View Coding (MVC) [16] using both preceding frames and the first view to generate the second view Simulcast (SIM) where both views of a stereo sequence are coded and transmitted independently.

3.5.1 Test Sequences

Table 4: Screenshots of test sequences

Screenshot	Characteristics / description
	Bullinger An anchorman presenting news, does not includes any camera movement
	Butterfly Computer generated animation, relatively small objects with motion
	Car Rear shot of a car on the road. Includes both camera and object movement, also exhibits global motion
	Horse Sequence of a horse eating grass.

Table 4 (continued)

Screenshot	Characteristics / description
	A&K Captured in studio environment with proper light, object motion with medium complexity without camera motion
	Lovebird Outdoor sequence without camera motion, includes relatively small object motion, detailed

3.6 Software Implementation

The source code is implemented by using C++ language. Borland C++ Builder v6 is used as the editor. Each metric is implemented with separate classes derived from the base class CMetric. Main advantages of this design pattern are easy to extend and scalability. In addition to classes designed specifically for metric calculation other algorithms and additional tools such as depth rendering, histogram segmentation, and video player have their own classes. OpenCV computer vision library 2.1.0 is used for image processing operations. Static structure of the framework drawn using UML notation is shown in Figure 3.11. Proposed metric is named with CLayerMetric and depends on CVSSIMMetric and CDepthUtil objects.

3.6.1 Framework Structure

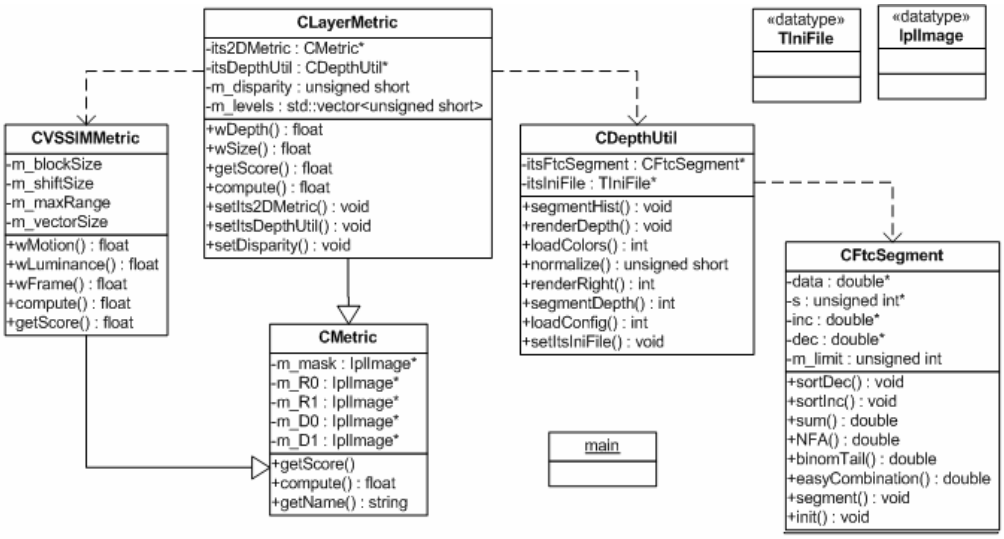

Figure 3.11: UML Diagram

3.6.2 Implemented Functions

Table 5: List of Implemented Functions

Class	Function Properties		
	visibility	return	name
CMetric	public	-	CMetric (CvSize size, int depth, int channels)
		float	compute()
		float	getScore()
		float	getWeight()
		string	getName()
		void	setR# (const IplImage *key, IplImage *mask = 0)
		void	setD# (const IplImage *key, IplImage *mask = 0)
		void	setMask (const IplImage *key)
		void	setCount (unsigned short key)
		void	setIndex (const unsigned short key)
CVSSIMMetric	private	float	wMotion()
		float	wLuminance (const CvRect rect)
		float	wFrame()
	public	-	CVSSIMMetric (const CvSize size, const int depth, const int channels)
		float	compute()
		float	getScore()
		void	setMaxRange(const CvSize key)

Table 5 (continued)

Class	Function Properties		
	visibility	return	name
CLayerMetric	private	float	wSize()
		float	wDepth (unsigned short zLayer, const unsigned short zNear)
	public	-	CLayerMetric (CvSize size, const int depth, const int channels)
		float	compute()
		void	setDisparity (unsigned short key)
		void	setItsDepthUtil (CDepthUtil *key)
		void	setIts2DMetric (CMetric *key)
		float	getScore()
CDepthUtil	private	void	segmentHist (const CvHistogram* hist, std::vector<unsigned short>* levels, const unsigned short disparity)
		unsigned short	normalize (const unsigned short level, unsigned short disparity)
		int	loadColors()
	public	int	renderRight (IplImage *src, IplImage *dest, IplImage *dep)
		int	segmentDepth (const IplImage *image, IplImage **imgHistogram, std::vector<unsigned short>* levels, const unsigned short disparity)

Table 5 (continued)

Class	Function Properties		
	visibility	return	name
CDepthUtil	public	int	loadConfig (long frame, vector<unsigned short>* levels)
		void	renderDepth (IplImage *left, IplImage *right, IplImage **dispLeft, IplImage **dispRight, unsigned short disparity)
		CvScalar	getColors (unsigned short index)
		void	setItsIniFile (TIniFile *key)
CFtcSegment	private	void	sortDec (int a, int b)
		void	sortInc (int a, int b)
		double	sum (double in[ ],unsigned int a, unsigned int b)
		double	NFA (unsigned int Lc, unsigned int Nc, double H, double Hc)
		double	binomTail (unsigned int n, unsigned int k, double p)
		double	easyCombination (unsigned int n, unsigned int k)
	public	void	setData (unsigned int x, double value)
		double	getData (const unsigned int x)
		void	setS (unsigned int x, unsigned int value)
		unsigned	getS (const unsigned int x)
		void	segment()
		void	init()

3.6.3 Graphical User Interface

Application design includes a graphical user interface (**GUI**) to support user-friendly operations. It is composed of a main form with additional input/output windows.

GUI has a 3 step selection procedure. Steps are as follows

- file selection
- metric selection
- output selection

In addition to these settings, color space and specific color components can also be selected.

Generic file types are supported such as BMP, JPEG and PNG for image; YUV and AVI for video files. For comparison and training purposes other metrics are also implemented. Complete list is given in Table 6. Executable snapshot is given in Figure 3.12.

Table 6: Included metrics in the software

Metric	3D extension
MSE	x
PSNR	x
SSIM	x
VSSIM	x
DELTA	x
MSAD	x
VSNR	x
SSA	√
DDM	√
LAYERS (proposed metric)	√

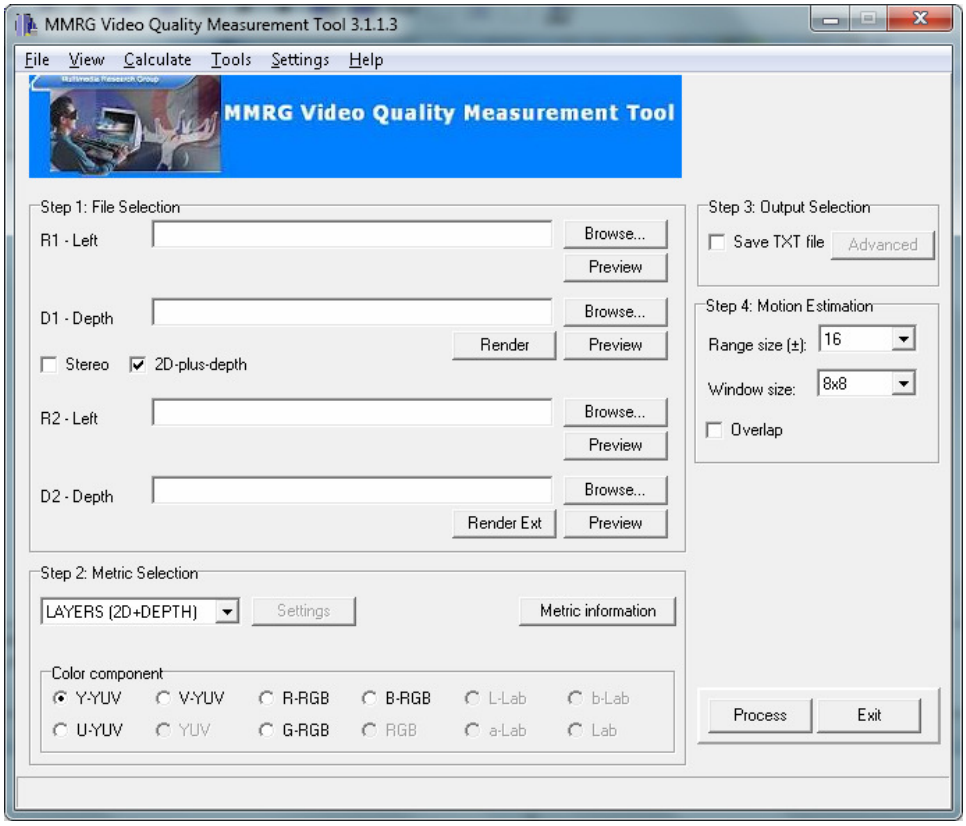

Figure 3.12: Graphical User Interface

Executable also includes a built-in video player, with customizable frame rate and color inversion capability. Screenshot taken while playing video is show in Figure 3.13. Using ‘Tools’ menu from menu bar, user can shrink, crop, split and resize video frame. Also stretch property can be changed from ‘Edit’ menu.

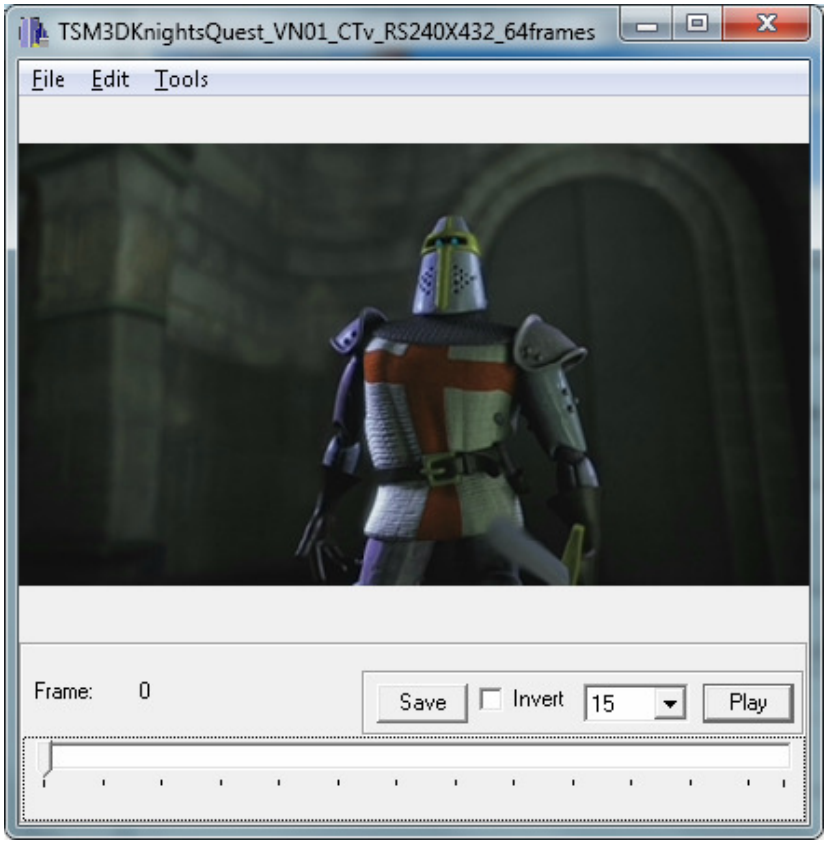

Figure 3.13: Player window

3.7 Simulation Results

Comparisons with other quality measures are given in this section with detailed graphs.

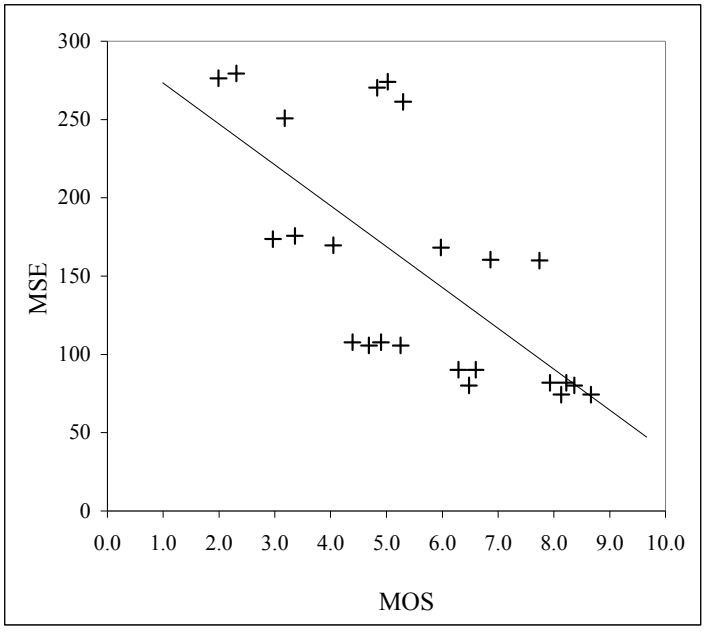

Figure 3.14: MOS vs. MSE

Results obtained by direct application of MSE metric to the test sequences are shown in Figure 3.14. More reliable results can be obtained by replacing MSE calculation with SSIM index but there exists some points outside the required region. Application of PSNR metric to individual layers can be seen in Figure 3.15 and Figure 3.16. Relative importances of layers are changed between theses two graphs. MOS vs. SSIM values can be seen in Figure 3.17. Application of VSSIM

creates better results but due to lack of weight adjustment for stereoscopic videos outlier ratio is still large in Figure 3.18.

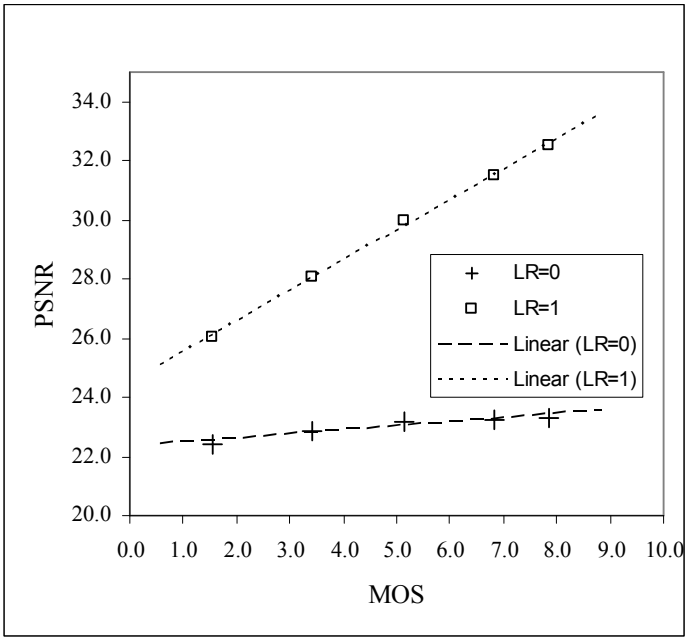

Figure 3.15: MOS vs. PSNR (Lovebird)

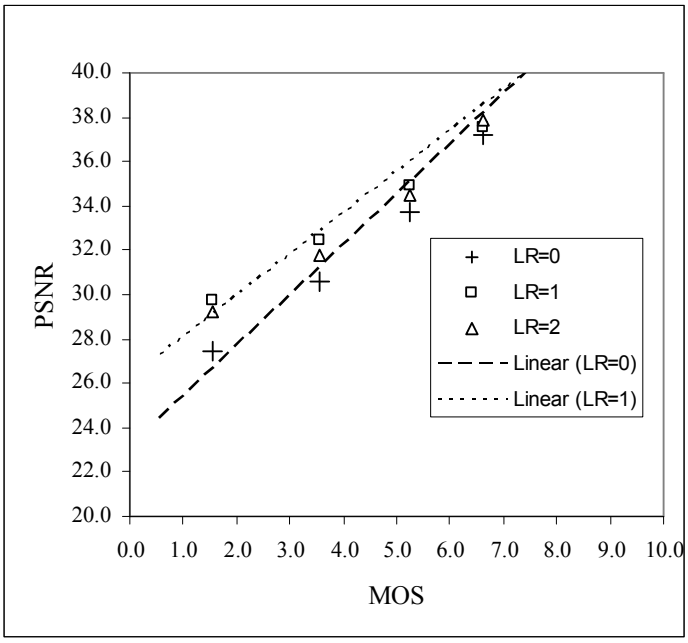

Figure 3.16: MOS vs. PSNR (A&K)

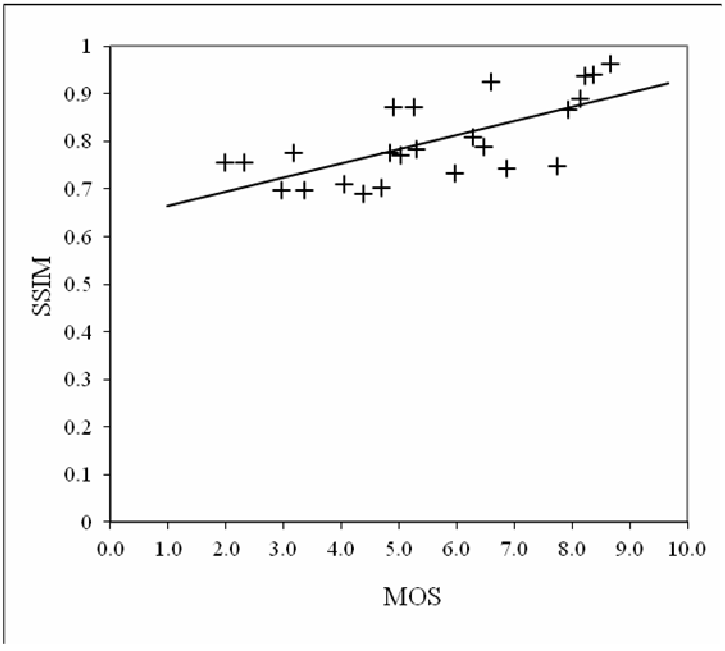

Figure 3.17: MOS vs. SSIM

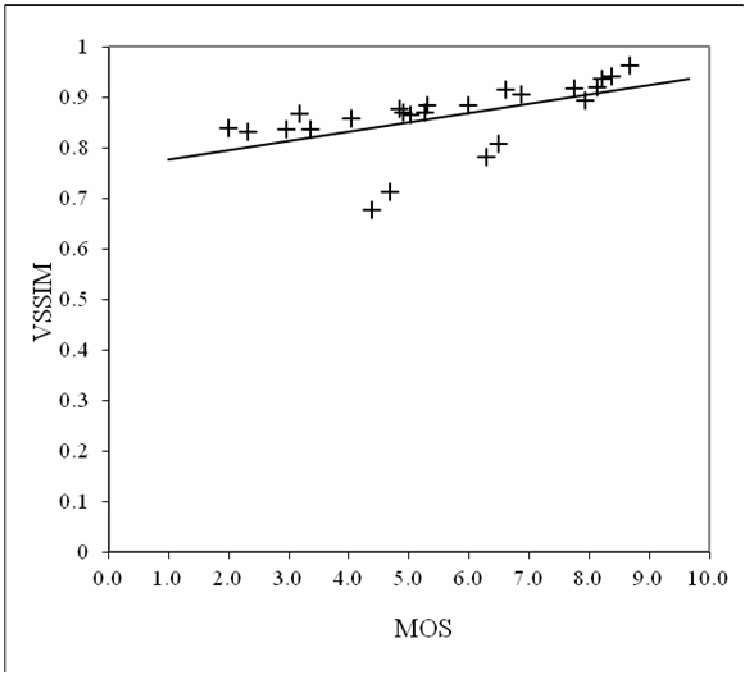

Figure 3.18 MOS vs. VSSIM

Results obtained from application of SSA 3D quality metric to the test sequences are show in Figure 3.19. There exist points outside the required region. SSA is also a feature-based metric with high computational efficiency.

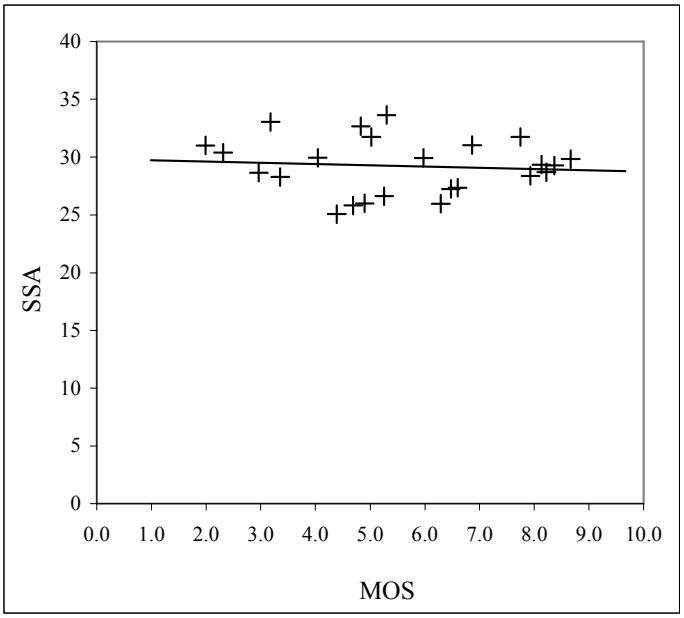

Figure 3.19 MOS vs. SSA

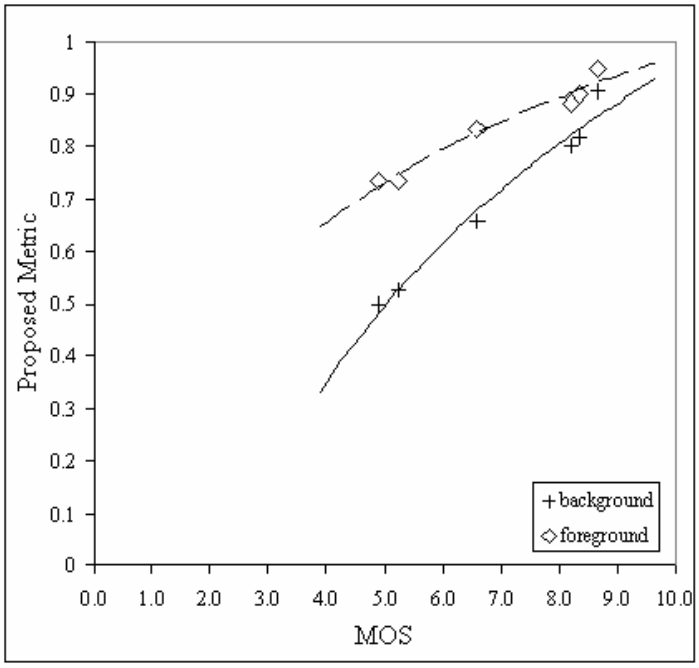

Figure 3.20: MOS vs. Proposed Metric (Butterfly)

In Figure 3.20 relation between individual layers of Butterfly sequence and MOS scores are given. It can be seen that correlation of background is greater than that of foreground.

In Figure 3.21 and Figure 3.22 results of Car and Bullinger sequences are given. Due to the content type properties of Bullinger test sequences it has shown different characteristics than the other set of videos. This situation is also observed during subjective experiments.

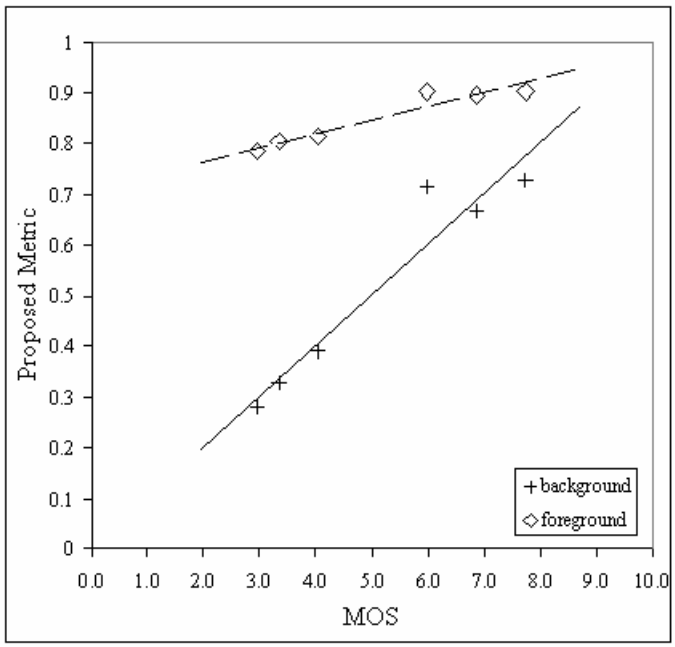

Figure 3.21: MOS vs. Proposed Metric (Car)

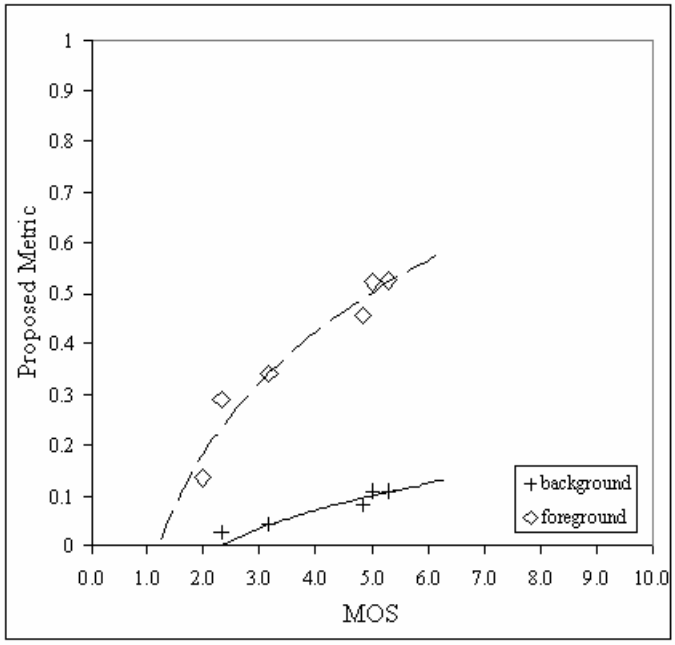

Figure 3.22: MOS vs. Proposed Metric (Bullinger)

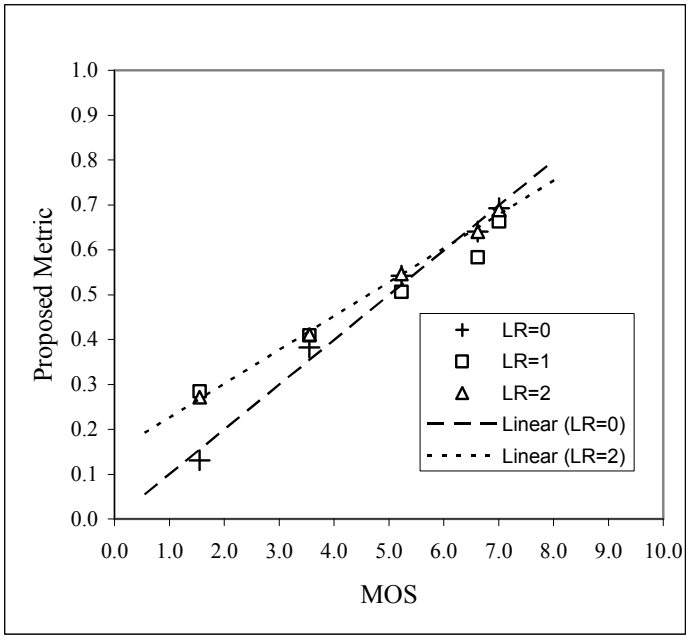

Figure 3.23: MOS vs. Proposed Metric (A&K)

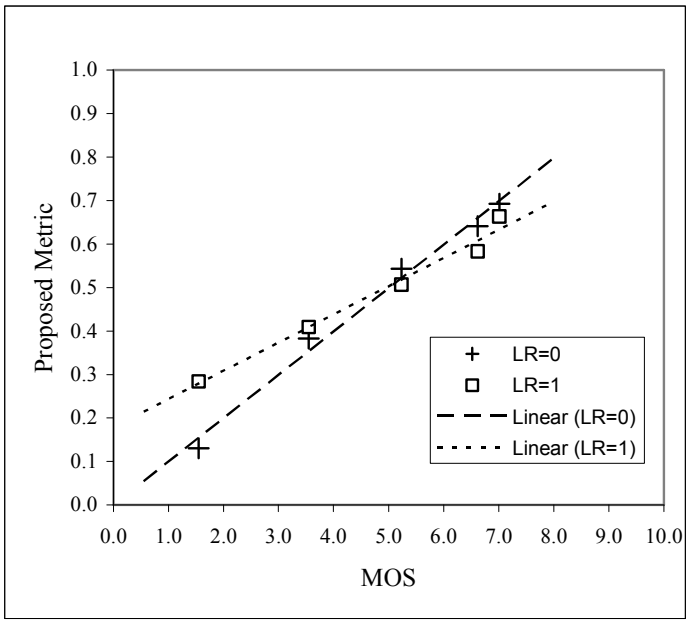

Figure 3.24: MOS vs. Proposed Metric (Lovebird)

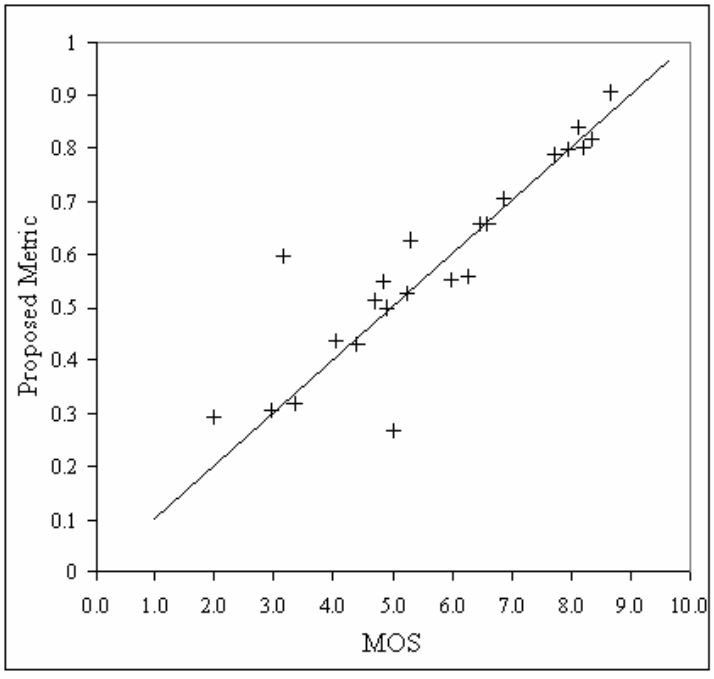

Figure 3.25: MOS vs. Proposed Metric (weighted)

With the weighting values properly trained, MOS vs. proposed metric distribution lies on the $x = y$ line.

Table 7: Performance comparison of proposed metric

Test Sequence	Spearman	Pearson
butterfly	0.005	0.990785
horse	0.004	0.969015
car	0.006	0.990241
bullinger	0.1	0.559903

CHAPTER 4

CONCLUSION

4.1 Summary

This study presents the effect of different depth layers with different disparity size to 3D perception. VSSIM index is used to estimate the quality because of its better performance in video. Proposed metric is given in Chapter 3. Related studies and basics of quality assessment are given in Chapter 2. A computer application using an object-oriented programming language is implemented for video processing and testing. Objective quality scores of test videos are presented and compared with MOS values.

4.2 Discussions

In our proposed method, VSSIM index includes both motion and luminance information, while histogram segmentation and depth layering measures 3D perception. These results demonstrate that overall 3D quality and the background layer quality are well correlated. This correlation indicates that background layer is more prone to coding errors. Differences in test results showed that content type plays an important role in the subjective evaluation of the 3D video. Main idea of the proposed metric is to divide depth information into segments where more features can be extracted. Then by extending a quality metric specifically designed for 2D video overall 3D quality can be estimated. Videos are processed in offline mode; an upgraded version can be made for real-time video processing.

4.3 Future Work

Further study can be made to find stereoscopic artifacts such as puppet theatre effect, cardboard effect and picket fence effect by extending proposed method. For this purpose in addition to histogram segmentation, object segmentation can be applied to 2D scenes. Also horizontal lines can be searched to find impurities in the perspective. In order to improve metric performance, scene changes can be marked and parameters can be revised accordingly.

In addition to quality evaluation, new methods in stereo content transmission can be improved, such as systems sending depth layers individually (layered 3D).

REFERENCES

[1] Huynh-Thu, Q., Le Callet, P., Barkowsky, M., “Video quality assessment: From 2D to 3D - Challenges and future trends”, ICIP10 (4025-4028).

[2] S.L.P. Yasakethu, D.V.S.X. De Silva, W.A.C. Fernando, and A. Kondoz, “Predicting sensation of depth in 3D video”, IET Electronic Letters, vol. 46, no. 12, pp. 837 – 839, Jun. 2010

[3] C. Hewage, M.G. Martini, "Reduced-reference quality metric for 3D depth map transmission", 3DTV Conference 2010, Tampere, Finland, 7-9 June 2010.

[4] Z. M. P. Sazzad, S. Yamanaka, et al., “Stereoscopic Image Quality Prediction”, International Workshop on Quality of Multimedia Experience, San Diego, CA, U.S.A., 2009.

[5] A. Mittal, A.K. Moorthy, J. Ghosh and A.C. Bovik, "Algorithmic assessment of 3D quality of experience for images and videos", IEEE Digital Signal Processing Workshop, Sedona, Arizona, January 04-07, 2011

[6] F. Speranza, W. J. Tam, R. Renaud, and N. Hur, "Effect of disparity and motion on visual comfort of stereoscopic images", Proceedings of SPIE 6055: 60550B (2006).

[7] P. Gorley and N. Holliman, “Stereoscopic image quality metrics and compression”, Proceedings of SPIE Stereoscopic Displays and Applications XIX, vol. 6803, pp. 680305, 2008.

[8] A. Benoit, P. Le Callet, et al., “Quality Assessment of Stereoscopic Images”, EURASIP Journal on Image and Video Processing, vol. 2008, 2008.

[9] G. Leon, H. Kalva, and B. Furht, “3D Video Quality Evaluation with Depth Quality Variations”, 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2008, pp. 301-304.

[10] De Silva, D.V.S.X.; Fernando, W.A.C.; Nur, G.; Ekmekcioglu, E.; Worrall, S.T.; , "3D video assessment with Just Noticeable Difference in Depth evaluation", Image Processing (ICIP), 2010 17th IEEE International Conference on , vol., no., pp.4013-4016, 26-29 Sept. 2010

[11] C. Hewage, S. Worrall, S. Dogan, S. Villette, and A. Kondoz, “Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video”, IEEE Journal of Selected Topics in Signal Processing, vol.3, 2009, pp. 304-318.

[12] Z. Wang, L. Lu and A.C. Bovik, "Video quality assessment based on structural distortion measurement", Signal Processing: Image Communication, Special issue on Objective video quality metrics, vol. 19, no. 2, February 2004.

[13] Junhwan Kim, Vladimir Kolmogorov, and Ramin Zabih, “Visual correspondence using energy minimization and mutual information”, In ICCV ’03: Proceedings of the Ninth IEEE International Conference on Computer Vision, page 1033, Washington, DC, USA, 2003. IEEE Computer Society.

[14] J. Delon, A. Desolneux, J-L. Lisani et A-B. Petro, “A non parametric approach for histogram segmentation”, IEEE Transactions on Image Processing, vol.16, no 1, pp.253-261, Jan. 2007

[15] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004

[16] ISO/IEC JTC1/SC29/WG11, Text of ISO/IEC 14496-10:200X/FDAM 1 Multiview Video Coding. Doc. N9978, Hannover, Germany (July 2008).

[17] P. Campisi, P. Le Callet, and E. Marini, “Stereoscopic images quality assessment”, in Proceedings of 15th European Signal Processing Conference (EUSIPCO '07), Poznan, Poland, September 2007.

[18] D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze, “Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception”, Advances in Multimedia, vol. 2010

[19] ISO/IEC JTC1/SC29/WG11, ISO/IEC CD 23002-3: Representation of auxiliary video and supplemental information. Doc. N8259, Klagenfurt, Austria (July 2007).

[20] A. Smolic, G. Tech and H. Brust, July 2010, “Report on generation of stereo video data base”, Mobile3DTV Project report, available online at http://mobile3dtv.eu/results/, last visited on September 2011

[21] D. Strohmeier, S. Jumisko-Pyykkö, K. Kunze, G. Tech, D. Bugdayci, and M. O. Bici, Feb. 2010, “Results of quality attributes of coding, transmission, and their combinations”, Mobile3DTV Project report

[22] Z. Wang and A. C. Bovik, “Mean squared error: love it or leave it? - A new look at signal fidelity measures,” IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98-117, Jan. 2009.

[23] D.M. Chandler, S.S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images", IEEE Transactions on Image Processing, Vol. 16 (9), pp. 2284-2298, 2007.

[24] H.R. Sheikh.and A.C. Bovik, "Image information and visual quality," IEEE Transactions on Image Processing, Vol.15, no.2, 2006, pp. 430-444.

[25] Digital Video Broadcasting (DVB): Transmission System for Handheld Terminals (DVB-H), ETSI EN 302 304 V1.1.1 (2004-11)

[26] K. Müller, P. Merkle, H. Schwarz, T. Hinz, A. Smolic, T. Wiegand, Multi-view Video Coding Based on H.264/AVC Using Hierarchical B-Frames, in Proc. PCS 2006, Picture Coding Symposium, Beijing, China, April 2006

[27] O. Schreer, P. Kauff, T. Sikora, eds, “3D video communication”, Wiley, 2005

[28] B. Backus, M. Banks, R. van Ee, J. Crowell, “Horizontal and vertical disparity, eye position, and stereoscopic slant perception”, Elsevier, 1999

[29] Howard, B. Rogers, “Binocular vision and stereopsis, 1995

[30] G. Desbordes, “Vision in the presence of fixational eye movements: insights from psychopsysics and neural modeling”, 2001

[31] G. Poggio, T. Poggio, “The analysis of stereopsis”, Annual Review of Neuroscience, 1984

[32] J. Koenderink, A. van Doorn, “Geometry of Binocular Vision and a Model for Stereopsis”, Springer, 1976

[33] W. IJsselsteijn, P. Seuntiëns, L. Meesters, “State-of-the-art in human factors and quality issues of stereoscopic broadcast television”, Advanced Three-dimensional Television System Technologies, ATTEST, 2002

[34] J. Gårding, J. Porrill, J. Mayhew and J. Frisby, “Stereopsis, vertical disparity and relief transformations”, Vision Research, Volume 35, Issue 5, March 1995, Pages 703-722

[35] W. IJsselsteijn, P. Seuntiens and L. Meesters, “Human factors of 3D displays”, in (Schreer, Kauff, Sikora, edts.) 3D Video Communication, Wiley, 2005.

[36] Cutting, P. Vishton, “Perceiving Layout and Knowing Distances: The Integration, Relative Potency, and Contextual Use of Different Information about Depth”, Academic Press, Inc, 1995, Perception of Space and Motion, p. 69

[37] M. Nawrot, “Depth from motion parallax scales with eye movement gain”, Journal of Vision, 2003, Vol. 3, No. 11, pp. 841-851

[38] S. Ohtsuka, S. Saida, “Depth Perception from Motion Parallax in the Peripheral Vision”, IEEE international Workshop on Robot and Human Communication, 1994

[39] A. Boev, D. Hollosi, A. Gotchev, “Classification of stereoscopic artefacts”, Mobile3DTV Project report, available online at http://mobile3dtv.eu/results/, last visited on September 2011

[40] M. McCauley and T. Sharkey, “Cybersickness: Perception of Self-Motion in Virtual Environments” in Presence: Teleoperators and Virtual Environments, 1(3), 311-318.,1992.

[41] E. Montag and M. Fairchild “Fundamentals of Human Vision and Vision Modeling”, in H. Wu and K. Rao, eds. “Digital video image quality and coding”, ch. 2, CRC press, 2006.

[42] D. Chandler, “Visual Perception (Introductory Notes for Media Theory Students)”, MSC portal site, University of Wales, Aberystwyth

[43] J. Yang, Ch. Hou, “Objective Quality Assessment Method of Stereo Images”, Proc. IEEE 3DTV Conference, Potsdam, Germany, May 2009

[44] H. Shao, X. Cao, G. Er, “Objective Quality Assessment of Depth Image Based Rendering In 3DTV System”, Proc. IEEE 3DTV Conference, Potsdam, Germany, May 2009

[45] M.H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Trans on Broadcasting, vol. 50, pp. 312-322, Sept. 2004

[46] S. Winkler, “Perceptual Video Quality Metrics – A review”, in H. Wu and K. Rao, eds. “Digital video image quality and coding”, ch. 5, CRC press, 2006.

[47] VQEG Objective test plan, VQEG, January 2000

[48] Final report from the Video Quality Expetrs Group on the validation of objective models of video quality assessment, VQEG, December 1999. Accessible at ftp://ftp.crc.ca/crc/vqeg, last visited on September 2011

[49] http://perso.telecom-paristech.fr/~delon/HistoSeg/, last visited on September 2011

[50] AYER, M., BRUNK, H., EWING, G., REID, W., AND SILVERMAN, E. An empirical distribution function for sampling with incomplete information. The Annals of Mathematical Statistics 26, 4 (1955), 641--647.

[51] BIRGÉ, L. The Grenander estimator: A nonasymptotic approach. The Annals of Statistics 17, 4 (1989), 1532--1549.

[52] ITU-R, “Methodology for the subjective assessment of the quality of television pictures,” Rec. BT.500, 2002.

[53] ITU-R, “Subjective assessment methods for image quality in high-definition television,” Rec. BT.710, 1998.

[54] ITU-R, “Methodology for the subjective assessment of video quality in multimedia applications,” Rec. BT.1788, 2007.

[55] ITU-T, “Subjective video quality assessment methods for multimedia applications,” Rec. P.910, 2008.

[56] ITU-R, “Subjective assessment of stereoscopic television pictures,” Rec. BT.1438, 2000.

[57] R.G. Kaptein, A. Kuijsters, M.T.M. Lambooij, W.A. IJsselsteijn, and I. Heynderickx, “Performance evaluation of 3D-TV systems,” in Proc. SPIE Image Quality and System Performance V, vol. 6808, Jan. 2008.

[58] M. Pölönen, T. Jarvenpaa, J. Hakksinen, “Comparison of near-to-eye displays: subjective experience and comfort,” J. Display Technol., vol. 6, no. 1, pp. 27–35, Jan. 2010.

[59] M.T.M Lambooij, W.A. IJsselsteijn, and I. Heynderickx, “Visual discomfort in stereoscopic displays: a review” in Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XIV, vol. 6490, Jan. 2007.

[60] Society for Information Display, “Display measurements standard”, http://icdm-sid.org/Public/DMS/ICDM-DMS.html, last visited on September 2011

[61] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, “New requirements of subjective video quality assessment methodologies for 3DTV,” in Proc. VPQM, Jan. 2010.

[62] M. Barkowsky, R. Cousseau, and P. Le Callet, “Influence of depth rendering on the quality of experience for an auto stereoscopic display,” in Proc. Int. Workshop QoMEX, pp. 192–197, July 2009.

[63] ITU-R, “Digital three-dimensional (3D) TV broadcasting,” Question ITU-R 128/6, 2008.

[64] ITU-T, “Objective and subjective methods for evaluating perceptual audiovisual quality in multimedia services within the terms of Study Group 9,” Question 12/9, 2009.

[65] A. Boev, M. Poikela, A. Gotchev, and A. Aksay, “Modeling of the stereoscopic hvs,” tech. rep., Mobile3DTV Project, July 2009.

[66] Sudipta N Sinha, "Graph Cut Algorithms in Vision, Graphics and Machine Learning", Integrative Paper, November, 2004, UNC Chapel Hill.

[67] GRENANDER, U. Abstract Inference. Wiley, New York, 1980.

[68] A. Boev, M. Poikela, A. Gotchev and A. Aksay, July 2010 “Modelling of the Stereoscopic HVS”, Mobile3DTV Project report, available online at http://mobile3dtv.eu/results/, last visited on September 2011