EVALUATION OF VISUAL QUALITY METRICS

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

BY

RAMAZAN FERHAT ÖLGÜN

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

SEPTEMBER 2011

Approval of the thesis:

EVALUATION OF VISUAL QUALITY METRICS

submitted by **RAMAZAN FERHAT ÖLGÜN** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and 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 : Ramazan Ferhat ÖLGÜN

Signature :

ABSTRACT

EVALUATION OF VISUAL QUALITY METRICS

Ölgün, Ramazan Ferhat

M.Sc., Departmant of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Gözde Bozdağı Akar

September 2011, 65 Pages

The aim of this study is to work on the visual quality metrics that are widely accepted in literature, to evaluate them on different distortion types and to give a comparison of overall performances in terms of prediction accuracy, monotonicity, consistency and complexity. The algorithms behind the quality metrics in literature and parameters used for quality metric performance evaluations are studied. This thesis also includes the explanation of Human Visual System, classification of visual quality metrics and subjective quality assessment methods. Experimental results that show the correlation between objective scores and human perception are taken to compare the eight widely accepted visual quality metrics.

Keywords: quality assessment, human visual system.

ÖZ

GÖRSEL KALİTE METRİKLERİNİN DEĞERLENDİRİLMESİ

Ölgün, Ramazan Ferhat

Yüksek Lisans., Departmant of Electrical and Electronics Engineering

Tez Yöneticisi: Prof. Dr. Gözde Bozdağı Akar

Eylül 2011, 65 Sayfa

Bu çalışmanın amacı literatürdeki kabul görmüş görsel kalite metrikleri üzerine çalışma yapmak, onları faklı tahrifatlarda değerlendirmek ve onları kestirim doğruluğunu, monotonluğunu, tutarlılığını ve kompleksliğini dikkate alarak karşılaştırmaktır. Literatürdeki kalite metriklerinin algoritmaları ve kalite metrik performanslarını değerlendirmek için kullanılan parametreler üzerine çalışma yapılmıştır. Araştırma ayrıca İnsan Görsel Sistemi, görsel kalite metrik sınıflandırması ve öznel kalite değerlendirme metodlarını içermektedir. Kabul görmüş sekiz farklı görsel kalite metriğini karşılaştırmak için nesnel sonuçlar ve insan algısı arasındaki ilişkiyi gösteren deneysel sonuçlar alınmıştır.

Anahtar Kelimeler: kalite yargısı, insan görsel sistemi.

ACKNOWLEDGMENTS

I would like to express my appreciation to my supervisor, Prof. Dr. Gözde Bozdağı Akar for her wisdom, advice and guidance. I would like to thank to Selim Sefa Sarıkan for his advice, my other colleagues in KAREL for their interest and my manager Dr. Alper Sarıkan for his support. I would also like to thank to TÜBİTAK – BİDEB for its financial support. Most of all, I would like to thank to my parents for their invaluable support.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

HVS	: human visual system
MOS	: mean opinion score
DMOS	: differential mean opinion score
SSIM	: structural similarity
PSNR	: peak signal to noise ratio
VIF	: visual information fidelity
VSNR	: visual signal to noise ratio
NQM	: noise quality measure
VSSIM	: video structure similarity
MSSIM	: multi-scale structural similarity
DSIS	: double-stimulus impairment scale
DSCQS	: double-stimulus continuous quality-scale
SSCQE	: single stimulus continuous qualty evaluation
SDSCE	: simultaneous double stimulus for continuous evaluation
CSF	: contrast sensitivity function
TID2008	: tampere image database 2008
VQEG	: video quality expert group
VQR	: visual quality rating
FR	: full reference
RR	: reduced reference
NR	: no reference

CHAPTER 1

INTRODUCTION

Digital imaging and video are increasingly used in wide variety of applications such as digital televisions, video conferencing, internet videos and so on. For all these applications, visual data may be corrupted during acquisition, compression, transmission, restoration and reproduction stages which degrades the quality observed by the end user. Methods of quantifying visual quality play an important role for the following applications:

- Dynamically monitoring and adjusting video processing applications.
- Optimizing algorithms of processing systems
- Benchmarking an image or video processing systems.

The obvious method of quantifying the visual quality is to get the opinion of human observers. However, subjective quality evaluations are impractical for real-time systems since they are inconvenient, expensive and time-consuming. The aim of the visual quality assessment research is to design quantitative measures for objective evaluation which is consistent with perceived subjective image quality.

Currently, most commonly used objective quality assessment approach is peak signal to noise ratio (PSNR) which is based on the intensity of distortion. Although it is computationally simple and widely used in video quality evaluation, it does not correlate well with the subjective evaluation [27]. A great deal of effort has been made to design objective quality assessment methods that are consistent with perceptual quality measures. One of the most popular methods is based on extracting structural information which is used to quantify the visual fidelity [10]. Its multi-scale and video extensions are also provided [16], [26]. The other method proposes to quantify the loss of image information to distortion process [11]. It approaches the image assessment as an information fidelity problem. Presenting a wavelet-based visual signal to noise ratio is another method to quantify the visual quality [13]. It operates by using both the low and mid level properties of Human Visual System (HVS). An extensive work has also been conducted by the Video Quality Expert Group (VQEG), established in 1997, to collect subjective ratings for a set of test sequences and to evaluate the performance of different objective video quality assessment systems with respect to these sequences. In Phase I tests [28], VQEG only achieved a limited success that only the subjective tests were successfully completed. Then, it continued its work on Phase II tests and still could not find an objective quality assessment metric with sufficient accuracy [25].

1.1 Scope of Thesis

The purpose of this work is to study the visual quality metrics that are widely accepted in literature, to evaluate them on different distortion types and to give a comparison of overall performances in terms of prediction accuracy, monotonicity, consistency and complexity. Knowing these performance evaluation criteria for each objective visual quality model, one can choose the most suitable one for a specific application to employ.

In Chapter 2, the HVS which forms the perception of the images and videos will be studied. In Chapter 3, after giving classification of the objective quality metrics, some of the widely used ones in literature are going to be studied. Finally, subjective quality evaluation methods will be described in this chapter. In Chapter 4, experimental results will be presented for the evaluation of the performance of objective quality metrics. Finally, we will make conclusions about study in Chapter 5.

CHAPTER 2

BACKGROUND INFORMATION

In order to achieve better understanding in designing quality assessment methods, it is necessary to take into account the HVS response to the image. In this chapter, anatomy of human eye, some important features of HVS and HVS modelling are studied.

2.1 Fundamentals of Human Vision and Vision Modelling

2.1.1 Eye

The human eye cross section is simply shown in Figure 2.1.1 [1]. The lens focuses the image on the retina surface and changes its shape under muscular control to perform proper focusing of near and distant objects. The iris controls the aparture of the lens and the amount of light entering the eye. The retina consists of an array of photoreceptors (cones and rodes). The cones are specialized to detect colors and function in bright illumination. The rods do not take part in color vision and function mainly in low lightining levels. The more sensitive cones are concentrated in a central region (the fovea) which means that high-resolution color vision is only achieved over a small area at the center of the field of view. Nerves connecting to the retina leave the eyeball through the optic nerve. The human brain processes and interprets visual information based on both received information and prior learned responses.

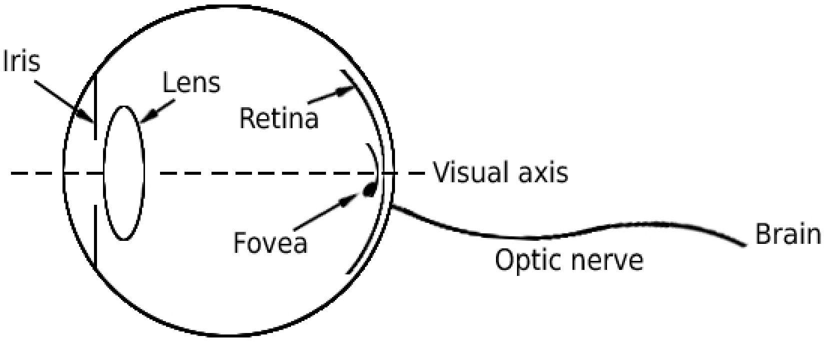


Figure 2.1.1: Human Eye

2.1.2 Color Vision

Appearance of an object's color results from the interaction of a light source, an object, and the visual system [2]. Early color perception of human is done in the retina, where the light-sensitive photoreceptors (cones) react to different wavelengths of light. Taking into account their ability to react short, medium and long wavelengths, these cones are called S, M and L-cones. Since these cones are sensitive to red, green and blue wavelenghts, the human vision is called trichromatic. In bright light the vision is called the photopic and the cone cells supply the color perception. In the low illumination levels the vision is scotopic and supplied by the rod cells.

2.1.3 Light Adaptation

The HVS is capable of adapting to a great range of light intensities [3]. There are three mechanisms for light adaptation to be distinguished in the HVS.

First is the mechanical variation of the pupillary aperture which is controlled by the iris. The iris reacts to differences in illumination level by varying pupil diameter which results in a 30-fold change of the quantity of light entering the eye. This adaptation process happens in a matter of seconds.

Second is the chemical processes in the photoreceptors which exist in both rods and cones. When the light is bright the concentration of photochemicals and their sensitivity reduces. In low light levels the production of photochemicals and thus the sensitivity increases. This adaptation process is rather slow that it takes up to an hour to adapt to a complete darkness.

Third is an adaptation at neural level. It involves the neurons in all layers of retina. By increasing or decreasing the signal output the neurons adapt to changing light intensities. This adaptation process is faster than the chemical adaptation in the photoreceptors.

2.1.4 Spatial Vision

The optics of the eye and the sampling of the visual scene by the retinal photoreceptors are two factors that need to be accounted for in the processing of spatial information [2]. Neural processing determines the visual response to spatial variation once the scene is sampled by the photoreceptors. The size and spacing of the retinal photoreceptors (rods and cones) determine the maximum spatial resolution [4].

Masking occurs when a stimulus cannot be detected because of the presence of another although it is visible by itself. Masking is strongest when the stimuli have similar characteristics, i.e. color, orientations or frequency. Spatial masking proves why similar artifacts are disturbing in certain regions of an image while they are hardly noticable elsewhere.

2.1.5 Temporal Vision

The sensitivity of the visual system to change over time and the perception of object motion are often linked together because the stimuli for the motion produce temporal variations in light intensity falling on the retina [2]. Movements of objects can be distinguished to rigid movements and form changes of the objects. The motion is detected as the transformation of spatial patterns of light entering the eye. Movement of all the images in the environment relative to the viewer creates a pattern of retinal motion named as optic flow. The object’s velocity perceived by the observer corresponds to its relative velocity to other objects in the visual field.

2.1.6 Visual Modelling

There are two approaches to model HVS, namely, neurobiological and psychophysical models [2]. The neurobiological models estimate the actual low-level process in the eye and optical nerve. Due to their overwhelming complexity, they are not useful in real-world applications.

The psychopyhsical models incorporate the aspects of human vision which are relevant to picture quality, such as color perception, contrast sensitivity and pattern masking. These models are based on psychopyhsical experiments and are implemented in a sequential process as shown in Figure 2.1.2.

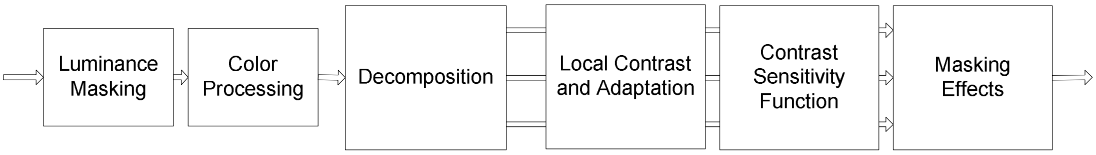


Figure 2.1.2: Block Diagram of a HVS Model

The color processing stage transforms the input signal into an acceptable perceptual color space. After this step the image is represented by one achromatic and two chromatic channels carrying color difference information. This stage also models the non-linear sensitivity of HVS to light, known as luminance masking [5].

HVS operates on multiple channels which are tuned to different spatial frequencies and orientations. This can be modeled by a multi-resolution filter bank or wavelet decomposition. It is believed that there are also channels tuned to different temporal frequencies.

The response of the HVS depends much on the local variations of luminance than on the absolute luminance which is a property known as Weber-Fechner law. Once the visual information is decomposed into channels this relative variation measure is used in vision model. The contrast perception also depends on the adaptation to a specific luminance or color level and local image content, which makes the precise modelling much more complex [6], [7], [8].

The decreasing sensitivity of HVS for higher spatial frequencies is one of the most important issues in HVS-modelling. This is typically parameterized by the contrast sensitivity function (CSF). A separate CSF for each channel of color space are modelled due to separate color and pattern sensitivity assumption.

Masking occurs when a stimulus cannot be detected because of the presence of another although it is visible by itself. The opposite effect of facilitation sometimes occurs if a stimulus can be detected due to the presence of another.

CHAPTER 3

QUALITY ASSESSMENT

In this chapter, objective and subjective quality assessment methods are studied. After a short explanation of classifications, some of the most widely used objective quality metrics are described. The subjective evaluation methods to evaluate the performance of objective quality metrics are also given.

3.1 Objective Quality Assessment

3.1.1 Classification of Objective Quality Metrics

Reference Information:

According to the information needed about the reference video, objective quality metrics can be divided into three categories [9]. The first category is the Full-Reference (FR) metrics which need the entire refence video in order to compare with test video frame-by-frame. The other type of metric is No-Reference (NR) metrics which only use video under test without the need of reference video. The third and the inbetween metric type in terms of the availability of reference information is Reduced-Reference (RR) metrics. They use some features extracted from the reference video and the comparison is done using those features.

Data Metrics:

Data metrics are based on byte-by-byte comparison between the reference and test videos. These type of metrics were designed to assess the fidelity of the data without considering the actual content. Data metrics are distortion-agnostic. The different types of distortions which are perceived differently by HVS would be rated the same by a data metric. Data metrics are also content-agnostic. The distortions affecting different parts of the image and are perceived differently by HVS would be rated the same by a data metric.

Although these type of metrics have no close relationship with the video quality perception of human observers, they are quite popular due to their computation simplicity.

Picture Metrics:

Picture metrics are used when the effects of distortion and content on visual quality perception wanted to be taken into account. There are two approaches in picture metrics category: the vision modelling and engineering approaches [9].

In vision modelling approach, metric design is based on modelling the HVS. They are incorporating the aspects of human vision such as color perception, contrast sensitivity and pattern masking obtained from psychophysical experiments.

In engineering approach, metric design is based on the extraction and analysis of certain features or artifacts. These can be the artifacts that are specifically introduced by such as encoders or structural elements such as contours. Then overall quality is estimated by looking for the strength of these features.

Packet- and Bitsream-based Metrics:

Packet- and Bitstream- based metrics are based on the parameters that can be obtained from the transported video stream over IP networks. Since these metrics need no or little decoding they require much lower processing requirements compared to the ones that examine fully decoded videos. Thus it is possible to assess more than one video parallely using these metrics.

3.1.2 PSNR

PSNR is a simple full-reference data metric that is widely used in video quality evaluation. PSNR as an engineering term is the ratio between the maximum possible power of a signal and the corrupting noise. In video quality evaluation, the signal is the original image data and the noise is the error introduced to the data in compression.

PSNR is derived by the mean squared error (MSE) which is formally given by

$$MSE = \frac{1}{n.m} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \| I(i,j) - \hat{I}(i,j) \|^2 \qquad (3.1.1)$$

where *I(i, j)* is the original frame at pixel position (i,j). *Î (i, j)* is the distorted frame at pixel position (i,j). m is the picture width and n is the picture height.

By setting the MSE in relation to the maximum pixel value in the frame which is 255 when the pixels are represented using 8-bits, PSNR is defined as

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right), [dB] \qquad (3.1.2)$$

The result gives an idea of how strong the noise effected the original data. If the two images are identical PSNR calculation is undefined due to division by zero, on the other hand if the image is compeletely distorted it has 0 value.

Although it is computationally simple and widely used in video quality evaluation, it does not correlate well with the subjective evaluation since it is distortion- and content- agnostic. Two videos with quantitatively the same PSNR values can have different subjective scores.

3.1.3 SSIM

SSIM (Structural Similarity) is a full-reference engineering approach which is popular for quality assessment of still images. SSIM is introduced to incorporate the structural information in quality assessment based on the assumption that HVS is highy sensitive to structural distortions [10].

The quality assessment system using SSIM is made up of easy to compute statistics of luminance comparison, contrast comparison and structure comparison as shown in Figure 3.1.1.

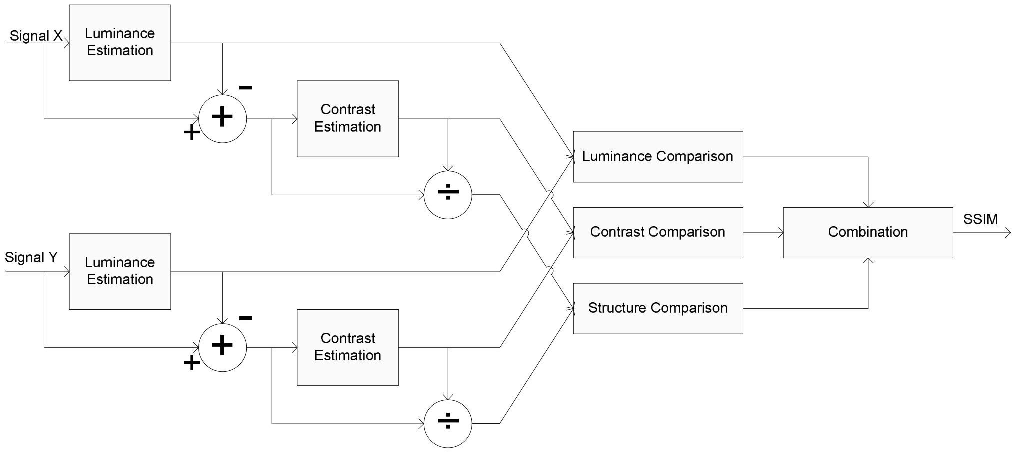


Figure 3.1.1 : Diagram of SSIM system
(adapted from [10])

X and Y are supposed to be the two nonnegative discrete image signals that have been aligned with each other and one of which is considered to be the perfect quality signal. For the luminance comparison of the image signals mean intensity estimation is used as a local statistic given by

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.1.3}$$

For the contrast comparison of the signals standard deviation estimation is used as in the below definition.

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{\frac{1}{2}} \tag{3.1.4}$$

For the structure comparison normalized signals of $\left(x-\mu_x\right)/\sigma_x$ and $\left(y-\mu_y\right)/\sigma_y$ are used.

Then the similarity measure is defined as

$$S(x,y) = f(l(x,y), c(x,y), s(x,y)) \tag{3.1.5}$$

where $l(x,y)$, $c(x,y)$ and $s(x,y)$ are comparison functions and $f(.)$ is the combination function needed to be defined. As stated in [10] similarity measure $S(x,y)$ has to satisfy the following conditions:

1) Symetry $S(x,y) = S(y,x)$

2) Boundedness: $S(x,y) \leq 1$

3) Unique maximum: $S(x,y) = 1$ if and only if (x,y) (in discrete represenatation $x_i = y_i$ for all i=1, 2, ..., N)

The luminance comparison function is defined as

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x{}^2 + \mu_y{}^2 + C_1} \tag{3.1.6}$$

where C_1 is added to avoid instability when $\mu_x{}^2 + \mu_y{}^2$ is very close to zero. In [10] it is choosen as

$$C_1 = (K_1 L)^2 \tag{3.1.7}$$

where L defined as the maximum pixel value, 255 if pixels are represented using 8-bits, and $K_1 << 1$ is a small constant.

The contrast comparison function is defined similarly as

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x{}^2+\sigma_y{}^2+C_2} \tag{3.1.8}$$

where $C_2=(K_2L)^2$ with $K_2<<1$ is added to avoid instability when $\sigma_x{}^2+\sigma_y{}^2$ is very close to zero.

The structure comparison function is defined as

$$s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \tag{3.1.9}$$

where C_3 is added similar to luminance and contrast measures. In addition correlation (σ_{xy}) is defined as

$$\sigma_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y) \tag{3.1.10}$$

Finally the combination of $l(x,y)$, $c(x,y)$ and $s(x,y)$ results in the SSIM index between the signals X and Y

$$SSIM(x,y)=[l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[s(x,y)]^{\gamma} \tag{3.1.11}$$

where $\alpha>0$, $\beta>0$ and $\gamma>0$ are parameters to adjust importance of the comparison functions. It is easy to show that comparison functions and the overal SSIM index all satisfies the three conditions given above. After simplifying the expression by setting $\alpha=\beta=\gamma=1$ and $C_3=\frac{C_2}{2}$ [10] it takes the form

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x{}^2+\mu_y{}^2+C_1)(\sigma_x{}^2+\sigma_y{}^2+C_2)} \tag{3.1.12}$$

While appliying SSIM index for image quality assessment, the local statistics μ_x, σ_x and σ_{xy} and SSIM index are calculated within a local 8x8 window over the entire image. In order to prevent blocking artifacts caused by this windowing approach, an 11x11 circular-symetric Gaussian weighting function is used. Then the local statistics μ_x, σ_x and σ_{xy} takes the below form

$$\mu_x=\sum_{i=1}^{N}w_ix_i \tag{3.1.13}$$

$$\sigma_x=\left(\sum_{i=1}^{N}w_i(x_i-\mu_x)^2\right)^{\frac{1}{2}} \tag{3.1.14}$$

$$\sigma_{xy}=\sum_{i=1}^{N}w_i(x_i-\mu_x)(y_i-\mu_y) \tag{3.1.15}$$

SSIM measure in this thesis uses the parameter values of $L = 255$, $K_1 = 0.01$ and $K_2 = 0.03$ where K_1 and K_2 are rather arbitrary and taken according to the [10]. Overall SSIM quality index of the entire image is calculated taking the mean SSIM.

$$MeanSSIM(X,Y) = \frac{1}{M}\sum_{j=1}^{M} SSIM(x_j, y_j) \qquad (3.1.16)$$

where X and Y are the original and the processed image signals; x_j and y_j are the image contents at jth local window; and M is the number of local windows of the entire image.

3.1.4 VIF

VIF (Visual Information Fidelity) is a full-reference vision modelling approach based on the relationship between image information and visual quality. VIF, as an image quality assessment method, combines the two quantities which are the information in the reference image and how much of this reference information can be derived from the test image.

In VIF measure proposed in [11], reference image is modeled as the output of a stochastic "natural" source which passes through the HVS channel and then processed by the brain. The information content of the original image is quantified as the mutual information between the input and output of HVS channel (between C and E). In the presence of an image distortion channel the same measure (between C and F) is quantified and the information of the test image extracted by the human brain is obtained. The Figure 3.1.2 shows the relation pictorially.

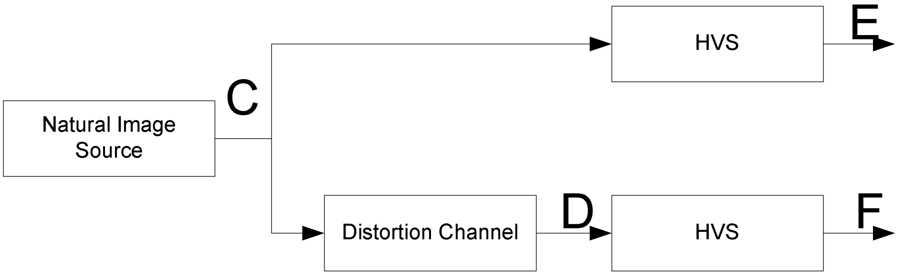


Figure 3.1.2 : Source, distortion and human visual system model relationship

The components of the proposed model are given below:

Source Model:

Natural scenes are classified as the images and videos that are captured using high quality capture devices operating in the visual spectrum. In VIF measure approach, natural images are modeled in the wavelet domain using Gaussian scale mixtures (GSMs). A GSM is a random field (RF) expressed as a product of two independent RFs [12]. Natural image source model output C is defined as,

$$C = S.U = \{S.\overrightarrow{U_i} : i \in I\} \tag{3.1.17}$$

where

$C = \{\overrightarrow{C_i} : i \in I\}$ I denotes the set of spatial indices for RF.

$S = \{S_i : i \in I\}$ is a RF of positive scalars.

$U = \{\overrightarrow{U_i} : i \in I\}$ is a Gaussian vector RF with zero mean and covariance C_U.

Distortion Model:

In real world, distortion types are approximated locally as a combination of blur and additive noise considering their perceptual annoyance. VIF measure also uses a signal attenuation and additive noise distortion model in the wavelet domain

$$D = GC + V = \{g_i.\overrightarrow{C_i} + \overrightarrow{V_i} : i \in I\} \tag{3.1.18}$$

where

$D = \{\overrightarrow{D_i} : i \in I\}$ is a RF from the subband in the test image.

$G = \{g_i : i \in I\}$ is a deterministic scalar gain field.

C is a RF from a subband in the reference image.

$V = \{\overrightarrow{V_i} : i \in I\}$ is a stationary additive zero-mean white Gaussian noise RF with variance $C_V = \sigma_V{}^2 I$.

Human Visual System Model:

In VIF metric, HVS model is described in the wavelet domain as a “distortion channel” that limits the information passing through it. The purpose of HVS model in VIF metric measurement is to quantifiy the uncertainity that HVS adds to the signal. This uncertainity is stated as an additive noise and modeled using stationary, zero mean, additive white Gaussian noise in the wavelet domain.

$$E = C + N \tag{3.1.19}$$

$$F = D + N^I \qquad (3.1.20)$$

where $N = \{\overrightarrow{N_i} : i \in I\}$ and $N^I = \{\overrightarrow{N^I{}_i} : i \in I\}$ are stationary RFs with covariance $C_N = C_{N^I} = \sigma_n{}^2 I$ where $\sigma_n{}^2$ is an HVS model parameter.

E and F are the visual signals at the output of HVS model from the reference and test images in one subband.

Visual Information Fidelity Criterion

After defining the source, distortion and HVS models, the VIF criterion can be derived. It is stated in [11] that visual quality relates well to the amount of information that human could perceive from the test image relative to the amount of information that human could perceive from the reference image. The information that could be ideally perceived by the subject from a particular subband in the reference and the test images are represented as $I(\overrightarrow{C}^N;\overrightarrow{E}^N \mid s^N)$ and $I(\overrightarrow{C}^N;\overrightarrow{F}^N \mid s^N)$. The mutual information $I(\overrightarrow{C}^N;\overrightarrow{E}^N \mid s^N)$ quantifies the amount of information that can be perceived by human when the reference image is being viewed. The s^N denotes a realization of S^N that could be thought as "model parameters" for a specific reference image to tune the natural scene model for that particular image. Incorporating multiple subbands the VIF measure is given by

$$VIF = \frac{\sum_{j \in subbands} I(\overrightarrow{C}^{N,j};\overrightarrow{F}^{N,j} \mid s^{N,j})}{\sum_{j \in subbands} I(\overrightarrow{C}^{N,j};\overrightarrow{E}^{N,j} \mid s^{N,j})} \qquad (3.1.21)$$

where it is summed over the subbands of interest and $\overrightarrow{C}^{N,j}$ represents N elements of the RF C_j which describes the coefficients from subband j.

There are some features obtained from equation 3.1.21. First, VIF is bounded below by zero. When all the information is lost in the distortion channel, VIF is calculated as zero ($I(\overrightarrow{C}^N;\overrightarrow{F}^N \mid s^N) = 0$). Second, when there is no information loss between reference and test image VIF is calculated as unity ($I(\overrightarrow{C}^N;\overrightarrow{F}^N \mid s^N) = I(\overrightarrow{C}^N;\overrightarrow{E}^N \mid s^N)$). Third, an enhancement of linear contrast of the original image results in a VIF value larger that unity. This feature makes VIF different from traditional QA methods by capturing the improvement in visual quality.

3.1.5 VSNR

VSNR (Visual Signal to Noise Ratio) is a full-reference vision modelling approach based on near-threshold and suprathreshold properties of HVS. VSNR is introduced to present a wavelet-based visual signal-to-noise ratio for quantifying the visual fidelity for natural images. This metric proposes two stages which use low-level and mid-level properties of HVS [13]. In the first stage, in order to determine if the distortions in the test image are visible, low-level HVS properties of contrast sensitivity and visual masking are used via wavelet based models. If the distortions in the test image are not visible, the test image is supposed to be of perfect visual fidelity. If the distortions are suprathreshold, as a second stage, the low-level HVS property of perceived contrast and mid-level HVS property of global precedence are used. These two HVS properties are then modeled and summed linearly to calculate VSNR metric.

Given the original image *I* and test image *Î,* the steps taken to calculate VSNR metric are given below in detail.

Preprocessing:

In order to take into account the viewing conditions and approximate the cortical decomposition performed by the HVS, a preprocessing process is performed. Below are the steps to be taken:

a) Compute the distortions contained in the test image via

$$E = \hat{I} - I \tag{3.1.22}$$

b) Perform M-level discrete wavelet transforms (DWTs) of I and E to obtain subbands $\{s_I\}$ and $\{s_E\}$.

c) Compute the vector of spatial frequencies $f = [f_1, f_2, ..., f_M]$ via

$$f_m = 2^{-m} rv \tan(\frac{\pi}{180}) \tag{3.1.23}$$

where $m = 1, 2, ..., M$, r is the display resolution in pixels per unit distance. v is the viewing distance expressed in the same unit.

Assessing the Detectability of the Distortion:

In order to determine if the distortions in the test image are visible, contrast thresholds are compared with the actual contrasts of the distortions for each fm in f. If the distortions are below the threshold, the test image $\hat{I}$ is supposed to be of perfect visual fidelity. Then no further analysis is required. Below are the steps to be taken:

a) For each f_m in f , compute the contrast detection threshold $CT(E_{f_m} \mid I)$ via

$$CT(E_f \mid I) = \frac{C(I_f)}{CSNR_f^{thr}} = \frac{C(I_f)}{a_0 f^{a_2 \ln(f)+a_1}} \tag{3.1.24}$$

b) For each f_m in f , measure the actual distortion contrast $C(E_{f_m})$ via

$$C(E_{f_m}) \approx \frac{k\gamma}{2^m \mu_{L(I)}(b+k\mu_I)^{1-\gamma}} \tag{3.1.25}$$
$$\times\sqrt{\sigma^2[s_{E(m,LH)}]+\sigma^2[s_{E(m,HL)}]+\sigma^2[s_{E(m,HH)}]}$$

where $\sigma^2[s_{E(m,\theta)}]$ denotes the standard deviation of the subband of E at mth level of decomposition with orientation $\theta = LH, HL,$ or HH .

c) If $C(E_{f_m}) < CT(E_{f_m} \mid I),\ \forall f_m \in f$, the test image $\hat{I}$ is supposed to be of perfect visual fidelity then no need to continue VSNR calculation ($VSNR = \infty$).

Computation of the Visual SNR:

In order to compute the finite VSNR metric following steps are taken:

a) Compute the perceived contrast of the distortions approximated by the total RMS distortion contrast [14] $d_{pc} = C(E)$ via

$$C(E) = \frac{1}{\mu_{L(I)}}\left(\frac{1}{N}\sum_{i=0}^{N}[L(E_i+\mu_I)-\mu_{L(E+\mu_I)}]^2\right)^{1/2} \tag{3.1.26}$$

where $L(P) = (b+kP)^\gamma$ is physical luminance. b represents black-level offset, k the pixel-to voltage-scaling factor and γ the gamma of the display monitor.

b) Compute the disruption of global precedence via

$$d_{gp} = \left(\sum_{m=1}^{M}[C^*(E_{f_m})-C(E_{f_m})]^2\right)^{1/2} \tag{3.1.27}$$

c) Compute VSNR

$$VSNR = 10\log_{10}(\frac{C^2(I)}{VD^2}) = 20\log_{10}(\frac{C(I)}{\alpha d_{pc} + (1-\alpha)\frac{d_{gp}}{\sqrt{2}}}) \qquad (3.1.28)$$

3.1.6 NQM

NQM (Noise Quality Measure) is a full reference vision modelling approach based on the degradation model. NQM models the degradation on image as a combination of linear frequency distortion and additive noise injection [15]. This metric proposes two complementary quality measures. First is the Distortion Measure (DM) for linear frequency distortion. Second is the NQM for the additive noise.

Distortion Measure (DM):

Distortion Measure is computed in three steps. First, the frequency distortion is found by comparing the model restored image and the restored image. Second, the deviation of this frequency distortion from allpass response of unity gain is computed. Third, the deviation is weighted by a lowpass CSF and integrated over the visible frequency range.

$$DM = \int_0^{f_{\max}} [1 - DTF(\frac{f_r}{f_N})]CSF(f_r)df_r \qquad (3.1.29)$$

where DTF (Distortion Transfer Function) is a model for the blurring in restoration algorithms. f_r is the radial frequency $f_r = \sqrt{{f_x}^2 + {f_y}^2}$ where f_x and f_y are the horizontal and vertical frequencies. f_N is the Nyquist frequency and $f_{\max}$ is the maximum radial frequency included in DM.

Noise Quality Measure (NQM):

NQM is computed in two steps. First, original and model restored images are separately processed to simulate the appearance of them to an observer through a contrast pyramid. The contrast pyramid by Peli [6] is used to measure

1) effects of distance, image dimesions and spatial frequency to the variation in contrast sensitivity
2) variation in the local luminance mean
3) contrast interaction between spatial frequencies
4) contrast masking effects

Second, the NQM is calculated by computing the SNR of model restored image and the restored image by

$$NQM(dB) = 10\log_{10}\left(\frac{\sum_x \sum_y O_s^2(x,y)}{\sum_x \sum_y (O_s(x,y) - I_s(x,y))^2}\right) \tag{3.1.30}$$

where $O_s(x,y)$ and $I_s(x,y)$ denote the simulated versions of the model restored image and the restored image, respectively.

3.1.7 MSSIM

MSSIM (Multi-Scale Structural Similarity) is a multi-scale extension of a SSIM metric explained in section 3.1.3. MSSIM [16] is introduced to incorporate the variations of viewing conditions to the previous single-scale SSIM measure. Display resolution and viewing distance are two conditions that are taken into account by moving to a multi-scale approach.

The quality assessment system using MSSIM is given in Figure 3.1.3.

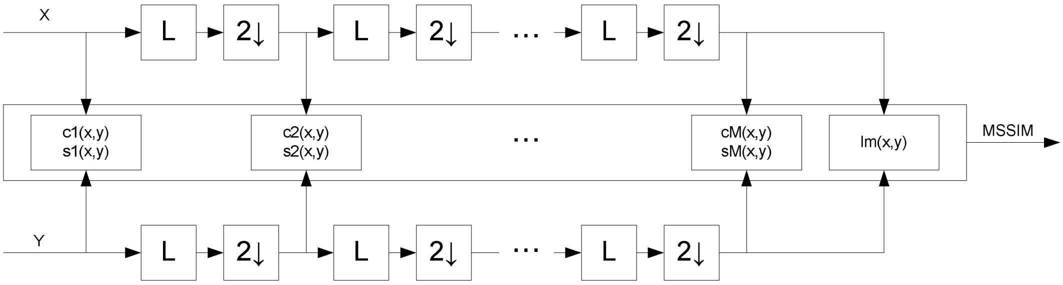


Figure 3.1.3 : Multi-scale structural similarity measurement system

The system takes the reference and distorted images and applies low-pass filter and downsamples the filtered images by 2 iteratively. The scaling is done from Scale 1 to Scale M and at the j-th scale the contrast and structure comparisons are calculated as $c_j(x,y)$ and $s_j(x,y)$. The limunance comparison $l_M(x,y)$ is calculated only at Scale M. By combining the measurements at different scales, an overall SSIM index is obtained as below

$$SSIM(x,y) = [l_M(x,y)]^{\alpha_M} . \prod_{j=1}^{M} [c_j(x,y)]^{\beta_j} . [s_j(x,y)]^{\gamma_j} \tag{3.1.31}$$

where α_M, β_j and γ_j are the parameters to adjust importance of the comparison functions.

For the simplicity of parameter selection it is taken as $\alpha_j = \beta_j = \gamma_j$ for all j's. In addition a normalization is done by $\sum_{j=1}^{M} \gamma_j = 1$. This normalization makes parameter settings comparable. In order to determine the relative values across different scales, an image synthesis approach is used. The resulting parameters obtained are $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$ and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$.

3.1.8 VSSIM

VSSIM (Video Structural Similarity) is a video extension of a SSIM metric explained in section 3.1.3. VSSIM is introduced to incorporate the temporal and spatio-temporal correlations between the adjacent frames of the videos [26]. In VSSIM, two methods are employed to weight the SSIM index. First, dark regions that do not usually attract fixations are used. Second, the regions in adjacent frames where large global motion occur are taken into account. The quality assessment system proposed by VSSIM is shown in Figure 3.1.4.

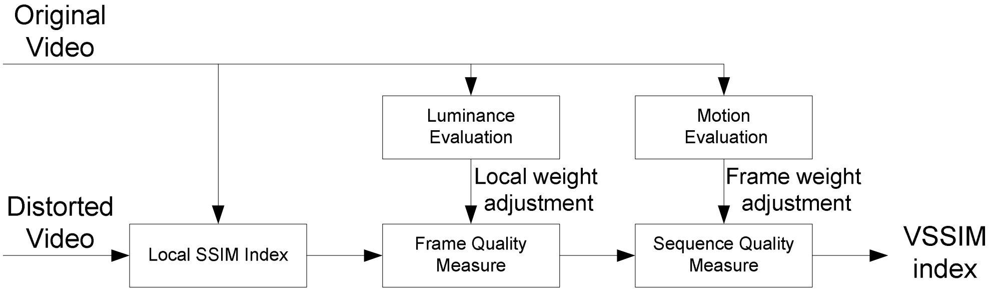


Figure 3.1.4 : Video quality assessment system

First, local sampling areas with sampling density of Rs per video frame are extracted. Then SSIM index is calculated for each sampling areas by

$$SSIM_{ij} = W_Y SSIM_{ij}^{Y} + W_{Cb} SSIM_{ij}^{Cb} + W_{Cr} SSIM_{ij}^{Cr} \qquad (3.1.32)$$

where $SSIM_{ij}^{Y}$, $SSIM_{ij}^{Cb}$ and $SSIM_{ij}^{Cr}$ denote the SSIM index of Y, Cb and Cr components of the j-th sampling window and i-th video frame. Weights are fixed and taken as $W_Y = 0.8$, $W_{Cb} = 0.1$ and $W_{Cr} = 0.1$.

Second, a frame-level quality index is calculated by combining the local quality values.

$$Q_i = \frac{\sum_{j=1}^{R_s} w_{ij} SSIM_{ij}}{\sum_{j=1}^{R_s} w_{ij}} \tag{3.1.33}$$

where Q_i is the quality index of the i-th frame and w_{ij} is the weighting value given to j-th sampling window in the i-th frame.

Third and the final overall quality of the video sequence is given by

$$Q = \frac{\sum_{i=1}^{F} W_i Q_i}{\sum_{i=1}^{F} W_i} \tag{3.1.34}$$

where F is the frame number and W_i is the weighting value given to i-th frame.

Local weighting adjustment method is based on the fact that dark regions in a frame of a video sequence usually do not draw much attention. These regions should be assigned smaller weighting values.

$$w_{ij} = \begin{cases} 0 & \mu_x \le 40 \\ (\mu_x - 40)/10 & 40 < \mu_x \le 50 \\ 1 & \mu_x > 50 \end{cases} \tag{3.1.35}$$

where μ_x is used as a local luminance estimation.

Frame weighing adjustment method is based on the fact that some type of distortions may not be as important in frames with large global motion. These frames should be assigned smaller weighting values.

$$W_i = \begin{cases} \sum_{j=1}^{R_s} w_{ij} & M_i \leq 0.8 \\ ((1.2 - M_i)/0.4)\sum_{j=1}^{R_s} w_{ij} & 0.8 < M_i \leq 1.2 \\ 0 & M_i > 1.2 \end{cases} \tag{3.1.36}$$

Where M_i is the motion level of the i-th frame and calculated as

$$M_i = \frac{(\sum_{j=1}^{R_s} m_{ij})/R_s}{K_M} \tag{3.1.37}$$

Where m_{ij} is the motion vector lenght of the j-th sampling window in the i-th frame. K_M is a normalization factor and taken as $K_M = 16$.

3.1.9 VQM

VQM (Video Quality Metric) is a general term given to all objective quality metrics evaluated in VQEG Phase II Full Reference Television tests [25]. In this thesis, only the General Model which was considered to be the most accurate among all is chosen to be used [31]. It was metric H in [25] and standardized by the American National Standards Institute (ANSI).

The General Model is a reduced-reference metric which utilizes features that are extracted from the original and processed video streams. The objective video quality measurement system is shown in Figure 3.1.5.

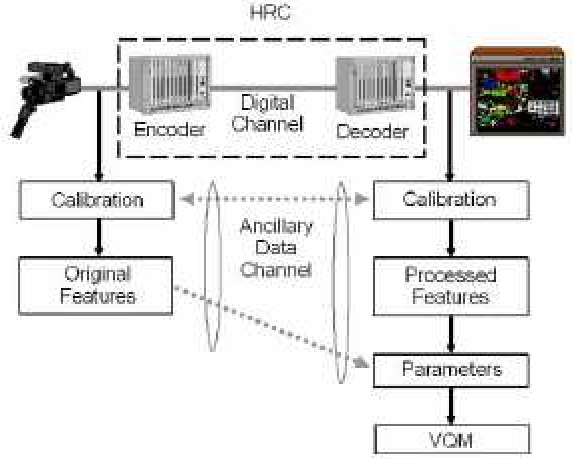


Figure 3.1.5 : VQM measurement system

The calibration of original and processed video streams includes spatial alignment, valid region estimation, gain & level offset calculation, and temporal alignment. The General Model measure is constructed by extracting perception-based features, computing video quality parameters, and combining parameters.

The quality feature is defined as a quantity of information extracted from a spatial-temporal sub-region of original and processed video streams. The quality parameters are indicative of perceptual changes in video quality that can be computed by comparing features extracted from the calibrated processed video with features extracted from original video. Seven independent General Model parameters are described as below.

- si_loss : The parameter detects a decrease or loss of spatial information (e.g. blurring).
- hv_loss : The parameter detects a shift of edges from horizontal & vertical orientation to diagonal orientation.
- hv_gain : The parameter detects a shift of edges from diagonal to horizontal & vertical.
- chroma_spread : The parameter detects changes in the spread of the distribution of two-dimensional color samples.
- si_gain : The parameter detects improvements to quality which result from edge sharpening or enhancement.
- ct_ati_gain : The parameter detects interactions between the amount of motion and spatial impairments. It also detects the interactions between the perceptibility of temporal impairments and the amount of spatial details.
- chroma_extreme : The parameter detects strong localized color impairments, such as those produced by digital transmission errors.

The General Model is constructed by linearly combining the video quality parameters as below:

$$\begin{aligned}\text{VQM} = & -0.2097 * \text{si\_loss} \\ & +0.5969 * \text{hv\_loss} \\ & +0.2483 * \text{hv\_gain} \\ & +0.0192 * \text{chroma\_spread} \\ & -2.3416 * \text{si\_gain} \\ & +0.0431 * \text{ct\_ati\_gain} \\ & +0.0076 * \text{chroma\_extreme}\end{aligned} \tag{3.1.38}$$

The General Model output values range from zero (no perceived impairment) to approximately one (maximum perceived impairment).

3.2 Subjective Quality Assessment

In this section subjective quality assessment methods to evaluate the video quality perceived by a human observer is presented. Recommendations [17] and [18] describe the methods and general features which is used to select from different options available those that best fits the objectives and circumstances of assesment problems.

The illimunation, viewing distance from observer to display and display properties are three factors that must be considered when conducting the subjective tests. Selection of test material is also an important issue to be considered. As stated in [18], in order to avoid boring the observers and to achieve a minimum reliability of the results, at least four different types of scenes should be chosen for the sequences. Observers' specifications are another important items in subjective quality assesments. As stated in [17], at least 15 observers should be used and they should be non-expert, in the sense that they are not directly concerned with television picture as part of their nominal work. Prior to subjective tests, the observers should be subjected to visual acuity and colour vision tests in order to produce reliable results. As stated in [17], a subjective assesment session should last up to half an hour. In addition, at the beginning of the first session , about five "dummy presentations" whose results are not be taken into account should be introduced to stabilize the observers' opinion. Observers should also be introduced adequately about the method before the subjective quality assessment session starts.

The subjective quality assessment methods proposed by the Recommendations [17] and [18] are depicted as below:

3.2.1 DSIS

The Double-Stimulus Impairment Scale (DSIS) method is used to check the degredation level of the test picture or sequence with respect to the reference picture or sequence. This method is also called as Degradation Category Rating (DCR) in [18]. The video sequences are presented in pairs: the first is the source reference and the second is the same source presented through the systems under test. Following the impaired sequence the observer is asked to vote on the second, keeping in mind the first. There are two variants to the structure of video sequences presentation:

- variant I : each video sequences are presented once
- variant II: each video sequences are presented twice

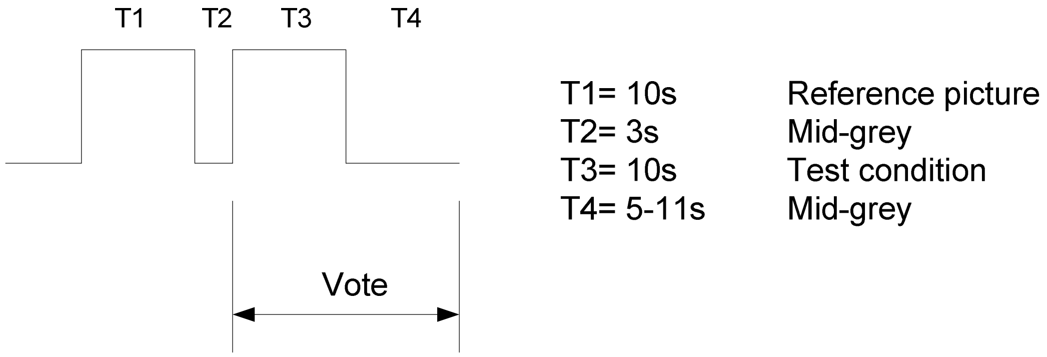


Figure 3.2.1 : Presentation structure of DSIS, variant 1

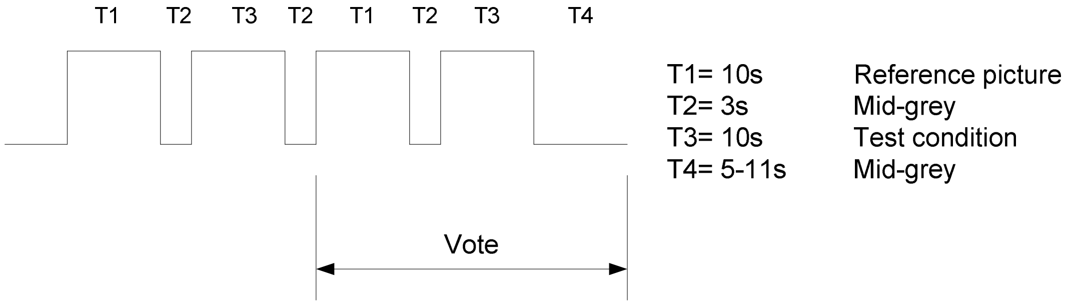


Figure 3.2.2 : Presentation structure of DSIS, variant 2

In order to rate the impairment of the test sequence five-level scale should be used:

- Imperceptible (5)
- Perceptible but not annoying (4)
- Slightly annoying (3)
- Annoying (2)
- Very annoying (1)

3.2.2 DSCQS

The Double-Stimulus Continuous Quality-Scale (DSCQS) method is especially useful when it is not possible to provide test stimulus test conditions that exhibit the full range of quality, stated in [17] . The video sequences are presented in pairs with randomized order of the reference and the test picture or sequences. Following the second sequence the observer is asked to vote on the both sequences. There are two variants to the structure of video sequences presentation:

- variant I : the observer, who is normally alone, is allowed to switch between the reference and test sequences until he establish his opinion of each.
- variant II: the multiple observers are shown the reference and test sequences twice to establish their opinion of each.

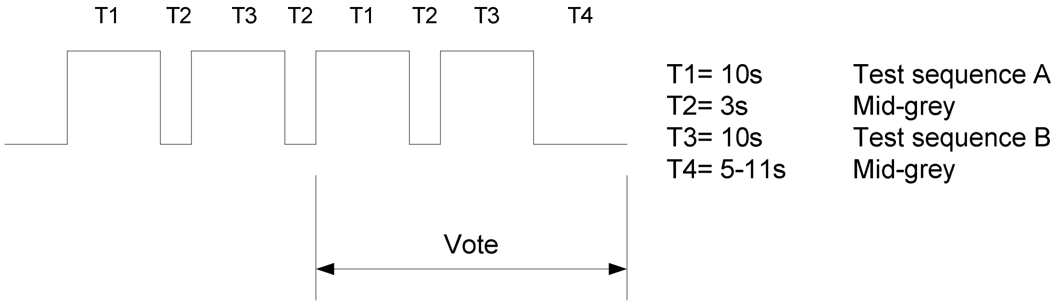


Figure 3.2.3 : Presentation structure of DSCQS

In order to rate the quality of both presentations the double vertical scale should be used. The scales are divided into five equal lenghts which correspond to quality scales as shown in the Figure 3.2.4. After the evaluation process, the pairs of assesments are converted to normalized scores in the range 0 to 100. Then the differences between normalized scores of reference and test sequences are calculated for each pair.

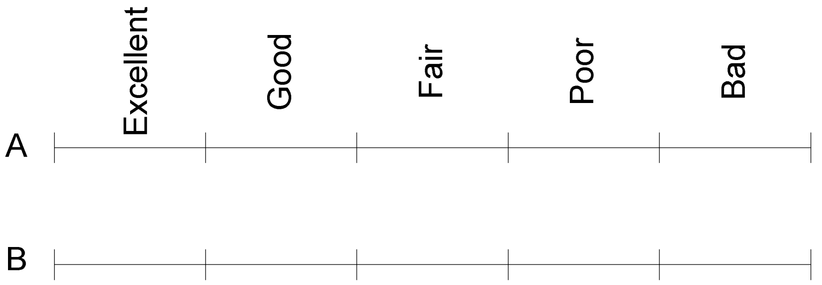


Figure 3.2.4 : Quality-rating form

3.2.3 SSCQE

The Single Stimulus Continuous Qualty Evaluation (SSCQE) method is useful to assess digitally coded video which have scene-dependent and time-varying impairments. In this method video sequences without a source reference are presented only once to the observer. Observers are continuosly assess the video sequence along the time on a linear scale by an electronic recording handset connected to a computer. Derivation of a single quality rating from the continuous quality results is currently under study as stated in [17].

3.2.4 SDSCE

The Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) method is suitable where fidelity of visual information affected by time-varying degradation has to be evaluated, stated in [17]. In this method video sequences are presented in pairs that reference and impaired sequences are displayed side by side in the same time. According to the [18], the subjects are asked to check the differences between the two sequences and to assess the fidelity of the video information along the time on a linear scale by an electronic recording handset connected to a computer. The subjects are aware of the reference and test sequences during assessment session. After the assessment session, data collected from the tests carried out can be processed to obtain a level of impairment.

CHAPTER 4

EXPERIMENTAL RESULTS

In this chapter quantitative performance evaluation of the objective quality metrics studied in Chapter 3 will be discussed. In order to evaluate the performance of objective quality metrics, it is necessary to obtain databases of test images or sequences from which the subjective quality scores have been experimentally collected. Two of the publicly-available databases downloaded and used during evaluations are Tampere Image Database 2008 (TID2008) and LIVE Video Quality Assessment Database. The reasons for the selection of these databases are due to their wide range of distortion types, variety of scenes and availability of subjective scores. In the subsequent sections, first, databases used during experiments are described. Second, the objective data analysis methods to evaluate the performance of objective quality metrics are discussed. Finally, the results of each objective quality metric are interpreted for each database.

4.1 Description of Databases

4.1.1 Tampere Image Database 2008

TID2008 is used to evaluate the FR objective image quality metrics [19], [20], [21]. TID2008 makes it possible to estimate a given metric correspondance to human perception.

The TID2008 contains a total of 1700 distorted images obtained from 25 reference images by applying 17 types of distortion with 4 level. All images are saved in database in 512x 384 24 bit format without any compression. Types of distortion with their correspondence to practical situation and accounted HVS properties are given in Table 4.1.1.

Table 4.1.1 : Tampere Image Database 2008 Distortion Types

No	Distortion	Correspondence to practical situation	Accounted HVS property
1	Additive Gaussian noise	Image acquisition	Adaptivity
2	Additive noise in color component is more intensive than additive noise in the luminance component	Image acquisition	Color sensitivity
3	Spatially correlated noise	Digital photography	Spatial frequency sensitivity
4	Masked noise	Image compression, watermarking	Local contrast sensitivity
5	High frequency noise	Image compression, watermarking	Spatial frequency sensitiviy
6	Impulse noise	Image acquisition	Robustness
7	Quantization noise	Image registration, gamma correction	Color, local contrast, spatial frequency
8	Gaussian blur	Image registration	Spatial frequency sensitivity
9	Image denoising	Image denoising	Spatial frequency, local contrast
10	JPEG compression	JPEG compression	Color, spatial frequency sensitivity
11	JPEG2000 compression	JPEG2000 compression	Spatial frequency sensitivity
12	JPEG transmission errors	Data transmission	Eccentricity
13	JPEG2000 transmission errors	Data transmission	Eccentricity
14	Non eccentricity pattern noise	Image compression, watermarking	Eccentricity
15	Local block-wise distortions	Inpainting, image acquisition	Evenness of distortion
16	Mean shift (intensity shift)	Image acquisition	Light level sensitivity
17	Contrast change	Image acquisition, gama correction	Light level, local contrast sensitivity

As seen in Table 4.1.1, TID2008 contains at least one distortion type for each HVS feature. White Gaussian noise was choosen as a model for additive zero-mean noise. Distortion type 2 was added in order to evaluate the objective metric correlation to HVS property to not equally perceive distortions in luminance and chrominance components. In order to test the metric correspondence to spatial frequency and local contrast sensitivity of HVS, spatially correlated noise, masked noise and high frequency noise were added as distortion types. These distortion types are typical for lossy image compression. Distortion type 6 is an impulse noise caused by coding/decoding errors in data transmission. The presence of this distortion might assist to evaluate the effectiveness of impulse noise removal methods. Image filtering is an important type of distortions for which it is crucial to have a suitable tool to assess visual quality of filtered image. In TID2008, a filter based on 3D Discrete Cosine Transform was used for filtering images corrupted by Gaussian noise. The images compressed by JPEG and JPEG2000 and decoded with errors in data transmission channels were also included in database. Such distortions are almost invisible since distorted fragments might occur to be similar to original texture and color of surrounding fragments. A specific type of distortion modeled by TID2008 group is non-eccentricity pattern noise. In this distortion type, a window of 15x15 pixels has been randomly taken from an original image and copied to few pixels nearby located fragment. Since it is difficult to identify such compact distortions, it shows that HVS is not sensitive to non-eccentric type of distortions. Another specific type of distortion was named as local block-wise distortion. It was modeled in such a way that 32x32 pixels of blocks with random colors were randomly copied to the different parts of image. Mean shift and contrast change of images were modeled as changes to two smaller and two larger values of the original image.

In TID2008, MOS was collected from the observers of three countries: Ukraine, Italy, and Finland. A total of 256428 comparisons of visual quality of distorted images have been performed by 838 observers.

Higher value of MOS (0 - minimum, 9 - maximum) corresponds to higher visual quality of the image.

4.1.2 LIVE Video Quality Assessment Database

LIVE Video Quality Database is intended to develop a database of videos that will be used to evaluate visual quality assessment methods [22], [23], [24].

The LIVE Video Quality Database is obtained from 10 uncompressed videos as reference videos with a wide variety of content. All video files have planar YUV 4:2:0 format and do not contain any headers. The spatial resolution of all videos is 768x432 pixels. The first 331776 bytes of each file correspond to the 8-bit Y component of the first frame, followed by 82944 bytes corresponding to the 8-bit U component of the first frame, followed by 82944 bytes corresponding to the 8-bit V component of the first frame. Frames are concatenated to form sequence files.

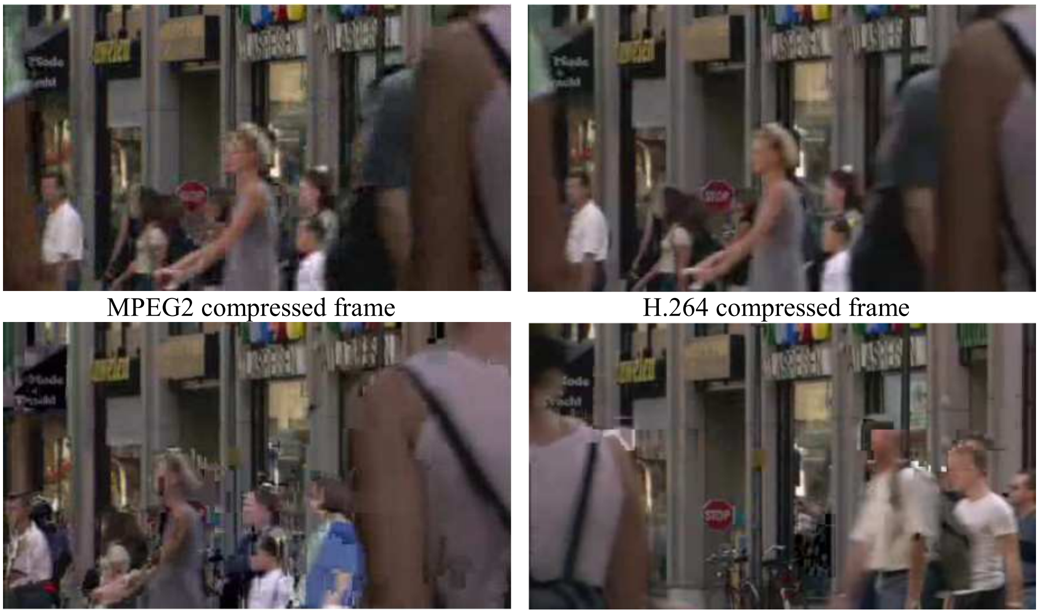
A total of 150 distorted videos were obtained from 10 reference videos using four different distortion types shown in Table 4.1.2.

Table 4.1.2 : LIVE Video Quality Assessment Database Distortion Types

No	Distortion	Correspondence to practical situation	Accounted HVS property
1	Wireless distortions	Data transmission	Eccentricity
2	IP distortions	Data transmission	Eccentricity
3	H.264 compression	H.264 compression	Spatial and temporal frequency sensitivity
4	MPEG2 compression	MPEG2 compression	Spatial and temporal frequency sensitivity

As seen in Table 4.1.2, LIVE Video Quality Database includes distortion types which are representative of present generation encoding and communication systems. IP and wireless distortions were created by simulating losses sustained by an H.264 compressed video over an IP network and wireless environment. The H.264 video streams were created using JM reference software and compression rates were choosen between 0.5-7 Mbps. For IP error, paterns supplied by the Video Coding Experts Group (VCEG), with loss rates of 3%, 5%, 10% and 20% were used. In Wireless type of distortion this error rates varied between 0.5-10%. Wireless channels are subjected to attenuation, shadowing and fading where a single bit error even causes a packet loss. Therefore, shorter packet sizes are used by encoding multiple slice per frame, where each packet contained a slice. In distortion type 3, H.264 compression was employed with compression rates varied from 200 Kbps to 5 Mbps. The MPEG2 video streams were created using reference software available from the International Organization for Standardization (ISO) and the compression rate was taken between 700 Kbps and 4 Mbps.

Each of the four distortion types in LIVE Video Quality Database is shown in Figure 4.1.1. The frames were taken from a video which was called “Pedestrian Area”. Notice that the blocking artifact in H.264 compressed frame is less than MPEG2. In addition, a wireless loss simulated frame exhibits errors that are restricted to smaller regions of frame than IP loss simulated.



MPEG2 compressed frame

H.264 compressed frame

IP loss simulated frame

Wireless loss simulated frame

Figure 4.1.1 : Frames of Pedestrian Area corrupted by each of the four distortion types in LIVE Video Quality Database

A single stimulus method is used to assess each video in the LIVE Video Quality Database. A total of 38 observers are asked to vote each video in a continuous quality scale. Difference Mean Opinion Scores (DMOS) which are collected from the subjects for original and distorted videos are available in the database.

$$DMOS = MOS_{reference} - MOS_{distorted} \qquad (4.1.1)$$

4.2 Objective Data Analysis

In order to provide quantitative performance evaluation of the objective quality metrics, methods provided in Video Quality Expert Group (VQEG) Phase II Final Report [25] are used. The outputs of objective quality metrics (the Video Quality Rating, VQR) and the observer Mean Opinion Scores (MOS) should be correlated in predictable and repeatable fashion. In order to remove any nonlinearity because of the subjective evaluation process and to compare the objective models in a common space, a nonlinear regression between the objective quality metrics and MOS values is used. The below logistic function is taken to fit to the data [MOS, VQR].

$$MOS_p = b_1 / (1 + \exp(-b_2(VQR - b_3))) \qquad (4.2.1)$$

This function is implemented by using MATLAB Curve Fitting Toolbox [29]. This nonlinear logistic function transforms the data set of VQR to a set of predicted MOS values. Then these predicted MOS values are compared with the MOS values taken from the subjective tests to evaluate objective metrics with respect to three aspects of their ability to estimate subjective assessments.

Prediction accuracy:

It is the ability to predict the subjective quality ratings with low error. The Pearson linear correlation coefficient between MOS and MOSp are calculated as a measure. It is obtained by dividing the covariance of two variables by the product of standard deviations as in Equation 4.2.2.

$$pearson_{X,Y} = corr(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma x \sigma y} \tag{4.2.2}$$

Root-mean-square error (RMSE) is an another common metric to measure the avarage error.

$$RMSE = \sqrt{E[(X-Y)^2]} \tag{4.2.3}$$

where X and Y are the paramaters that the correlation between them are to be calculated. For the objective quality analysis these are the MOS and MOSp values.

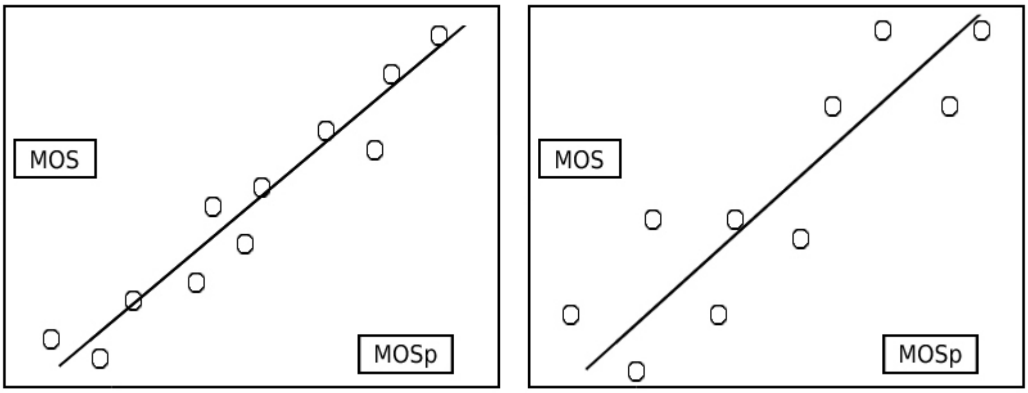


Figure 4.2.1 : Models showing the higher accuracy (left image) and lower accuracy (right image)

Prediction monotonicity:

It is the degree to which the model’s predictions agree with the relative magnitudes of subjective quality ratings. The Spearman correlation coefficient between MOS and MOSp are calculated as a measure. Spearman correlation is defined as the Pearson correlation between the ranked variables.

$$spearman_{X,Y} = corr(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad (4.2.4)$$

where X and Y are the paramaters that the correlation between them are to be calculated. For the objective quality analysis these are the MOS and MOSp values.

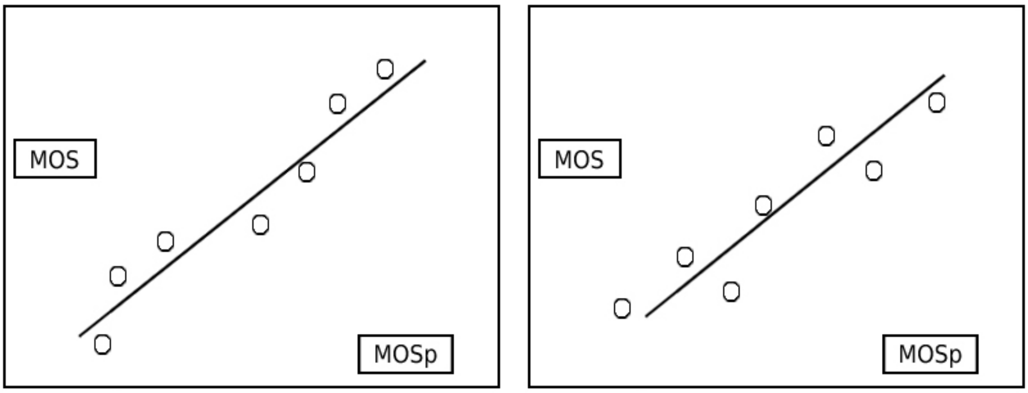


Figure 4.2.2 : Models with more monotonicity (left image) and less monotonicity (right image)

Prediction consistency:

It is the degree to which the model maintains prediction accuracy over the range of test sequences. The Outlier ratio is used to quantify the prediction consistency.

$$\text{Outlier Ratio} = (\text{total number of outliers})/\text{N} \qquad (4.2.5)$$

where outlier point is a point for which

$$|Qerror[i]| > 2 \times \text{MOS\_Standard\_Error} \qquad (4.2.6)$$

where $Qerror$ is the difference between the MOS and MOS_p values for a given objective quality score and twice the MOS Standard Error is used as a threshold for defining outliers.

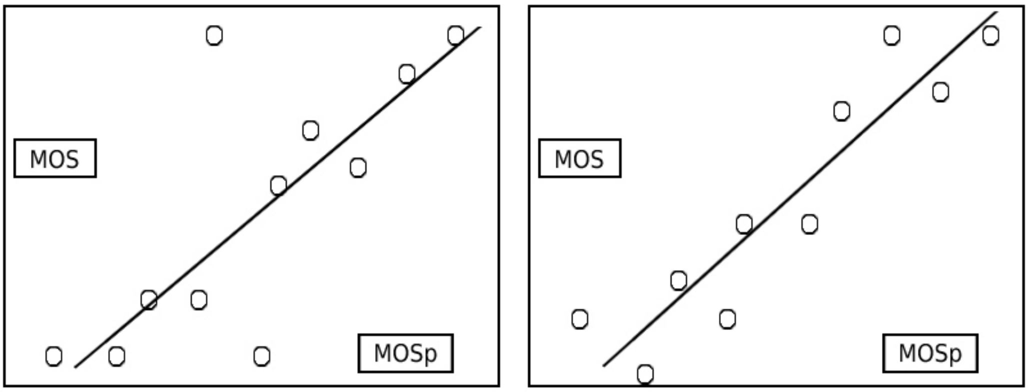


Figure 4.2.3 : Models with large outlier ratio (left image) and small outlier ratio (right image)

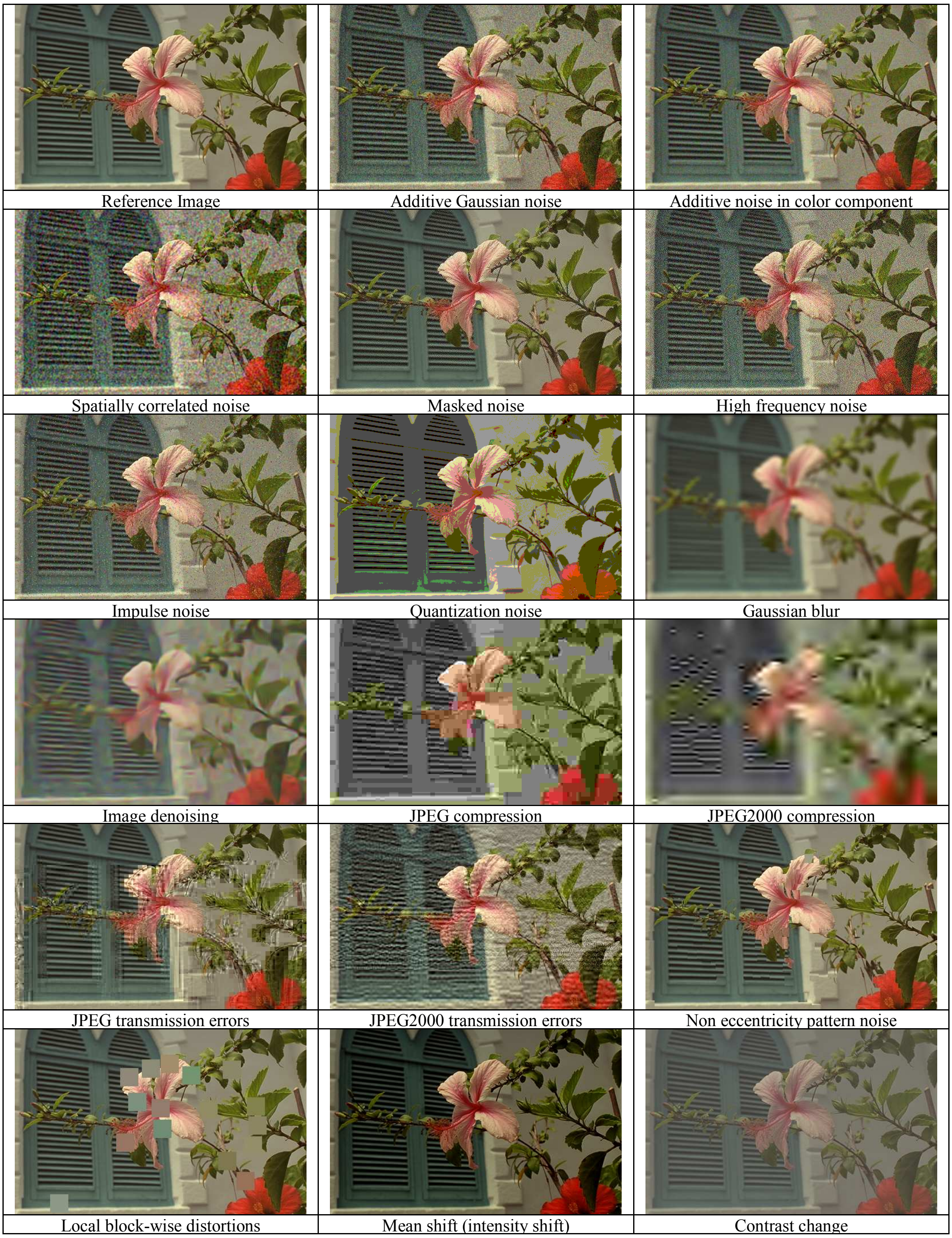
4.3 Performance of Objective Quality Metrics

In order to evaluate the performances of objective quality metrics, outputs of each metric are calculated separately for the impaired images and videos. Then, the correlation of these output values with the HVS are examined. All the metrics' outputs except VSSIM are calculated by using Metrix MUX Matlab package (see Appendix-A). The VSSIM metric is implemented by using C++ language.

For example, Table 4.3.1 presents the objective quality metrics' outputs for each distortion types of an image in TID2008 which is shown in Figure 4.3.1.

Table 4.3.1 : Objective quality metrics' outputs of images in Figure 4.3.1

Distortion	PSNR	SSIM	VIF	VSNR	NQM	MSSIM
Additive Gaussian noise	24.4	0.58	0.39	21.1	23.3	0.91
Additive noise in color component	30.4	0.79	0.61	28.9	28.5	0.96
Spatially correlated noise	24.4	0.60	0.25	13.1	14.2	0.82
Masked noise	24.2	0.77	0.68	31.2	27.5	0.97
High frequency noise	18.5	0.34	0.34	18.8	21.0	0.85
Impulse noise	24.4	0.61	0.38	18.1	20.3	0.91
Quantization noise	23.9	0.77	0.33	14.9	11.5	0.89
Gaussian blur	22.6	0.55	0.13	10.3	9.60	0.77
Image denoising	23.7	0.70	0.17	12.1	10.7	0.83
JPEG compression	24.9	0.73	0.18	14.6	14.1	0.88
JPEG2000 compression	21.6	0.52	0.04	8.84	7.09	0.66
JPEG transmission errors	20.9	0.50	0.19	8.83	9.39	0.73
JPEG2000 transmission errors	21.2	0.46	0.16	8.31	8.75	0.69
Non eccentricity pattern noise	25.3	0.90	0.54	12.4	11.1	0.92
Local block-wise distortions	24.7	0.92	0.75	10.8	8.7	0.88
Mean shift (intensity shift)	16.2	0.84	0.88	34.3	21.8	0.98
Contrast change	22.1	0.86	0.64	10.2	6.34	0.84



Reference Image

Additive Gaussian noise

Additive noise in color component

Spatially correlated noise

Masked noise

High frequency noise

Impulse noise

Quantization noise

Gaussian blur

Image denoising

JPEG compression

JPEG2000 compression

JPEG transmission errors

JPEG2000 transmission errors

Non eccentricity pattern noise

Local block-wise distortions

Mean shift (intensity shift)

Contrast change

Figure 4.3.1 : Distortion Types in Tampere Image Database 2008

4.3.1 PSNR

Table 4.3.2 : Performance of PSNR on Tampere Image Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Additive Gaussian noise	0.9345	0.9129	0.01	0.2208
Additive noise in color component	0.9241	0.9060	0.03	0.1985
Spatially correlated noise	0.9526	0.9190	0	0.1934
Masked noise	0.8729	0.8489	0.04	0.2952
High frequency noise	0.9714	0.9298	0	0.2312
Impulse noise	0.8578	0.8726	0.06	0.2672
Quantization noise	0.8760	0.8709	0.02	0.4048
Gaussian blur	0.8569	0.8698	0.03	0.6142
Image denoising	0.9462	0.9421	0	0.5263
JPEG compression	0.8684	0.8725	0.02	0.8575
JPEG2000 compression	0.8643	0.8139	0.01	1.897
JPEG transmission errors	0.7576	0.7509	0.06	0.8649
JPEG2000 transmission errors	0.8539	0.8322	0.05	0.4253
Non eccentricity pattern noise	0.5859	0.5813	0.22	0.8587
Local block-wise distortions	0.6383	0.6199	0.13	0.5177
Mean shift (intensity shift)	0.7084	0.6973	0.10	0.4124
Contrast change	0.6033	0.5875	0.11	0.9908
All Images	0.5274	0.5531	0.17	1.141

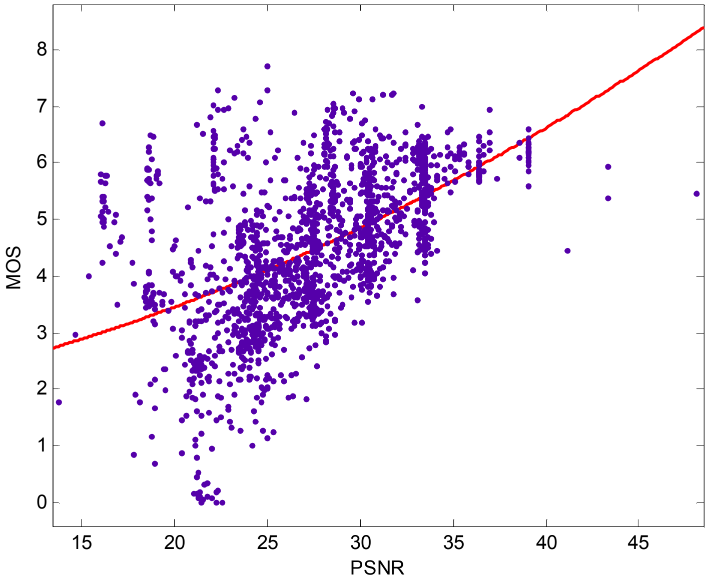


Figure 4.3.2 : Scatter plot of MOS versus PSNR on Tampere Image Database (b1=23.04, b2=0.19, b3=7.25)

Table 4.3.2 contains PSNR results for all images in TID2008 classified by the type of distortion. As shown in table, PSNR has poor overall performance (pearson correlation factor for all images is 0.5274). When considering the distortion types seperately, the correlation values increases. Since PSNR assumes that the distortion is caused by additive signal-independent noise, it performs better for such additive noise based distortions. As can be seen, distortion type of non-eccentricity pattern noise gives the lowest correlation with human perception. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.22. Therefore it can be stated that the results are consistent for all types of distortions. Figure 4.3.2 shows the relationship between MOS and PSNR pictorially in a scatter plot.

Table 4.3.3 : Performance of PSNR on LIVE Video Database for each distortion type

Distortion	**Pearson Corr.**	**Spearman Corr.**	**Outlier Ratio**	**RMSE**
Wireless	0.6574	0.6204	0	8.095
IP	0.4853	0.4718	0.1	8.644
H.264	0.5439	0.4729	0.2	10.27
MPEG2	0.4006	0.3831	0.15	9.083
All Videos	0.5433	0.5233	0.0666	9.433

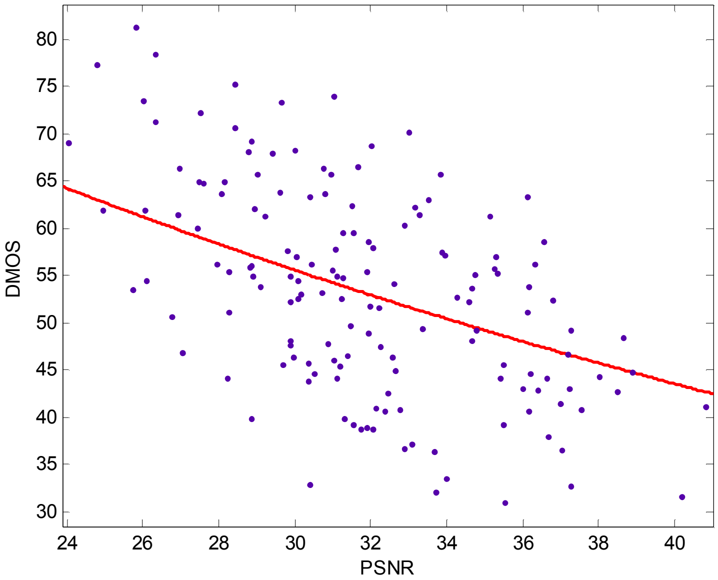


Figure 4.3.3 : Scatter plot of DMOS versus PSNR on LIVE Video Database (b1=1022, b2=-0.089 , b3=-32.56)

Table 4.3.3 contains the PSNR results for all videos in LIVE Video Database classified by the distortion type. As in the TID2008, PSNR performs weak also in this database (pearson correlation factor for all videos is 0.5433). Considering PSNR model correlation with human perception for each distortion types, the best correlation is obtained in wireless distortion. On the contrary, the worst correlation is in MPEG2 compression type of distortion. For the prediction consistency, outlier ratio for all types of distortion is between 0 and 0.2. It can be concluded that the results are consistent. Figure 4.3.3 shows the relation of subjective scores with PSNR for all the videos in LIVE Video database as a visual illustration. Since subjective scores are given in DMOS, the higher the PSNR values the lower the DMOS.

4.3.2 SSIM

Table 4.3.4 : Performance of SSIM on Tampere Image Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Additive Gaussian noise	0.7676	0.7891	0.0100	0.3995
Additive noise in color component	0.7905	0.7929	0	0.3185
Spatially correlated noise	0.7934	0.8129	0	0.405
Masked noise	0.8005	0.8043	0	0.3639
High frequency noise	0.8483	0.8389	0	0.5149
Impulse noise	0.6819	0.7161	0.07	0.3845
Quantization noise	0.7542	0.7988	0.07	0.5445
Gaussian blur	0.9327	0.9388	0	0.433
Image denoising	0.9382	0.9393	0	0.563
JPEG compression	0.9114	0.8930	0	0.7119
JPEG2000 compression	0.9263	0.9278	0	0.7475
JPEG transmission errors	0.8419	0.8402	0	0.9856
JPEG2000 transmission errors	0.8231	0.8407	0.01	0.4729
Non eccentricity pattern noise	0.6772	0.6957	0.02	0.7774
Local block-wise distortions	0.8861	0.8808	0.19	0.5169
Mean shift (intensity shift)	0.6918	0.7168	0.01	0.4225
Contrast change	0.4968	0.5134	0.01	1.079
All Images	0.6449	0.6459	0.0188	1.028

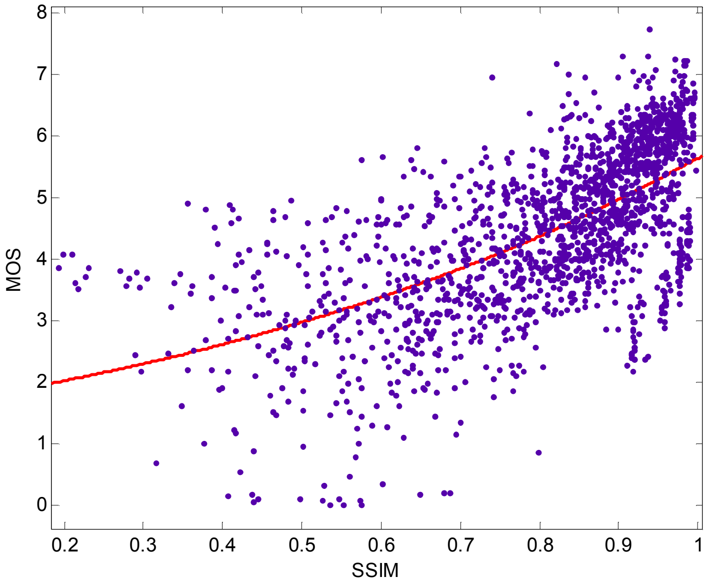


Figure 4.3.4 : Scatter plot of MOS versus SSIM on Tampere Image Database (b1=368.7, b2=0.207, b3=21.26)

Results in Table 4.3.4 show the performance of SSIM metric on TID2008 for each distortion type. As it can be observed from this table, SSIM has fair overall performance (pearson correlation factor is 0.6449). The better correlations between SSIM and subjective results are obtained in image denoising and gaussian blur distortion types. The worst performance to predict the subjective quality ratings is in contrast change distortion. Since SSIM is sensitive to distortions that corrupt spatial correlation such as block compression, blur and noise and insensitive to contrast and mean changes, the results are not suprising. In terms of consistency criteria, outlier ratio is between 0 and 0.19, and it can be concluded that the results are consistent. Figure 4.3.4 shows the relationship between MOS and SSIM pictorially in a scatter plot.

Table 4.3.5 : Performancce of SSIM on LIVE Video Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Wireless	0.5277	0.5221	0.05	9.114
IP	0.5369	0.4700	0	8.324
H.264	0.6705	0.6561	0.025	8.374
MPEG2	0.3759	0.5583	0.3	8.107
All Videos	0.5413	0.5251	0.0466	9.324

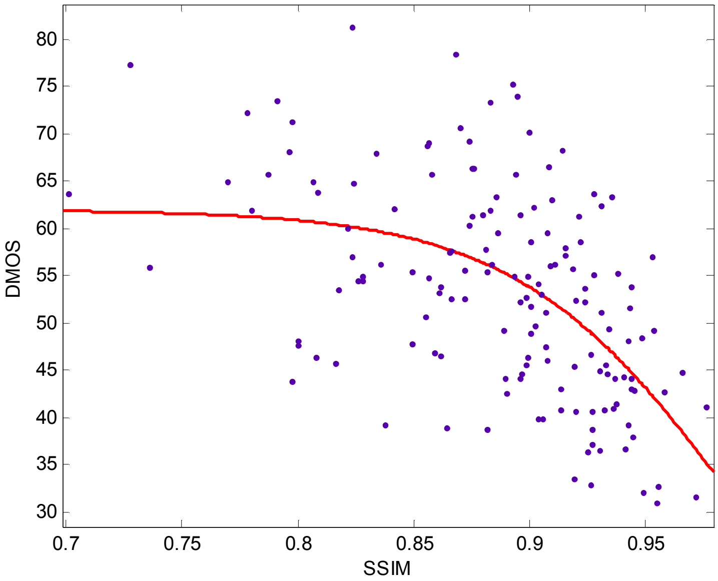


Figure 4.3.5 : Scatter plot of DMOS versus SSIM on LIVE Video Database (b1=61.94, b2=-1.104, b3=1.953)

Table 4.3.5 presents the results of SSIM metric for LIVE video database. As seen in the table, SSIM performs poor in LIVE video database (pearson correlation factor is 0.5413). Regarding the distortion types, SSIM is more correlated with subjective scores in H.264 than MPEG2 compression type of distortions. Possible explanation of this result is the reduced blockiness in H.264 due to deblocking filtering. Since SSIM is sensitive to block compression artifacts, it is not suprising to have better correlation for H.264 type of distortion. In addition, SSIM performs almost the same in IP and wireless distortion types. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.3, thus the results can be stated as consistent. Figure 4.3.5 shows the relationship between DMOS and SSIM pictorially in a scatter plot.

4.3.3 VIF

Table 4.3.6 : Performance of VIF on Tampere Image Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Additive Gaussian noise	0.8674	0.8799	0	0.3087
Additive noise in color component	0.8955	0.8785	0	0.2312
Spatially correlated noise	0.8616	0.8702	0	0.3226
Masked noise	0.8922	0.8698	0	0.2732
High frequency noise	0.9457	0.9075	0	0.3163
Impulse noise	0.8161	0.8331	0	0.3004
Quantization noise	0.7893	0.7956	0.01	0.5155
Gaussian blur	0.9384	0.9546	0	0.4118
Image denoising	0.9296	0.9189	0	0.5995
JPEG compression	0.9538	0.9170	0	0.5197
JPEG2000 compression	0.9579	0.9713	0	0.573
JPEG transmission errors	0.8760	0.8582	0	0.6389
JPEG2000 transmission errors	0.8352	0.8509	0.01	0.4494
Non eccentricity pattern noise	0.7441	0.7608	0	0.7079
Local block-wise distortions	0.8403	0.8320	0	0.3663
Mean shift (intensity shift)	0.5936	0.5132	0.14	0.5022
Contrast change	0.8869	0.8190	0	0.5738
All Images	0.7982	0.7496	0.0159	0.8092

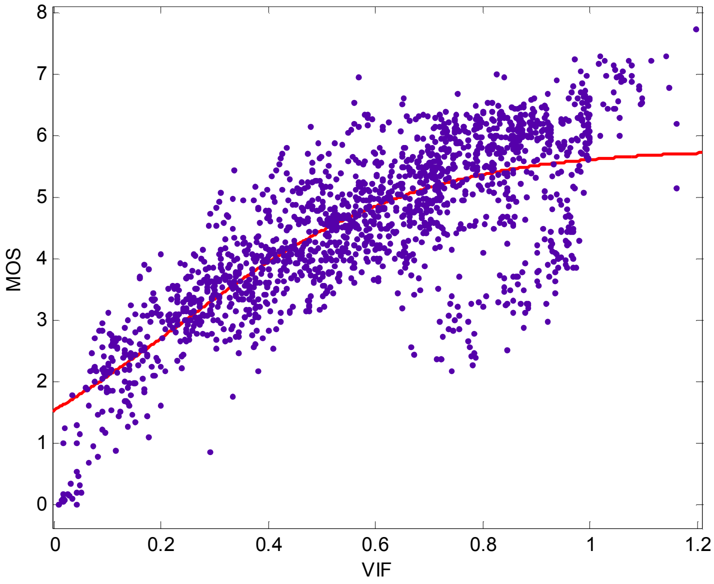


Figure 4.3.6 : Scatter plot of MOS versus VIF on Tampere Image Database (b1=5.788, b2=4.435, b3=0.229)

Table 4.3.6 contains VIF results for all images in TID2008 classified by the type of distortion. As shown in table, VIF has good overall performance (pearson correlation factor for all images is 0.7982). Considering the distortion types seperately, correlation values increases. Since VIF uses a signal attenuation and additive noise distortion model, it performs better for such distortions that the model should be able to synthesize images whose perceptual annoyance is close to actual distortion. As can be seen, VIF performs good in JPEG and JPEG2000 compressions, gaussian blur and image denoising types of distortion that can successfully be modelled. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.14. Therefore it can be stated that the results are consistent. Figure 4.3.6 shows the relationship between MOS and VIF pictorially in a scatter plot.

Table 4.3.7 : Performance of VIF on LIVE Video Database for each distortion type

Distortion	**Pearson Corr.**	**Spearman Corr.**	**Outlier Ratio**	**RMSE**
Wireless	0.5517	0.5317	0.05	8.949
IP	0.5917	0.5506	0	7.97
H.264	0.6497	0.6349	0.025	8.58
MPEG2	0.6299	0.6331	0.25	8.959
All Videos	0.5541	0.5541	0.02	9.231

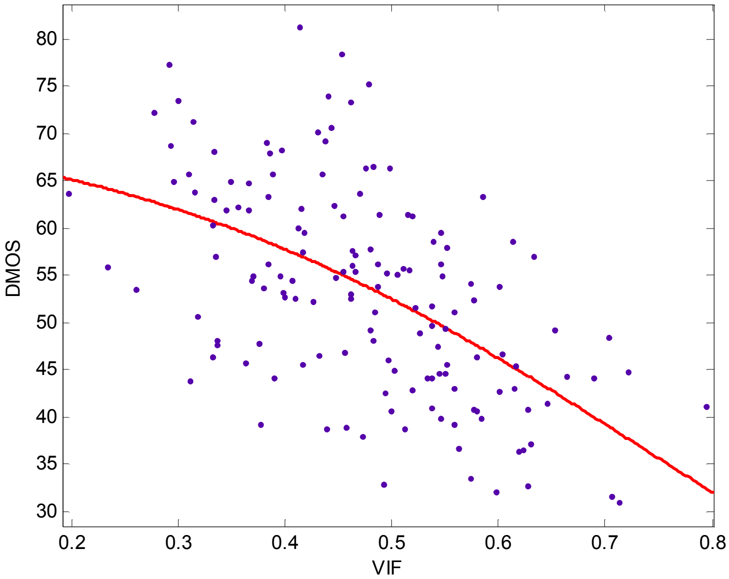


Figure 4.3.7 : Scatter plot of DMOS versus VIF on LIVE Video Database
(b1=72.77, b2=-0.43, b3=2.38)

Table 4.3.7 presents the results of VIF metric for LIVE video database. As seen in the table, VIF performs poor in LIVE video database (pearson correlation factor is 0.5541). Regarding the distortion types, VIF is more correlated with subjective scores in H.264 than MPEG2 compression type of distortions. In addition, VIF performs better in IP distortion than in wireless distortion type. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.25, thus the results can be stated as consistent. Figure 4.3.7 shows the relationship between DMOS and VIF pictorially in a scatter plot.

4.3.4 VSNR

Table 4.3.8 : Performance of VSNR on Tampere Image Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Additive Gaussian noise	0.7495	0.7278	0.11	0.4106
Additive noise in color component	0.7761	0.7793	0.06	0.3276
Spatially correlated noise	0.7557	0.7665	0.09	0.4162
Masked noise	0.7543	0.7295	0.11	0.3972
High frequency noise	0.8902	0.8811	0.01	0.4435
Impulse noise	0.6268	0.6472	0.23	0.4052
Quantization noise	0.8158	0.8269	0.07	0.4855
Gaussian blur	0.9246	0.9330	0.02	0.4538
Image denoising	0.9423	0.9286	0	0.5447
JPEG compression	0.9345	0.9174	0	0.6158
JPEG2000 compression	0.9467	0.9515	0	0.6411
JPEG transmission errors	0.8085	0.8055	0.04	0.7799
JPEG2000 transmission errors	0.7682	0.7909	0.06	0.5232
Non eccentricity pattern noise	0.5727	0.5716	0.14	0.8687
Local block-wise distortions	0.2748	0.1926	0.36	0.657
Mean shift (intensity shift)	0.3760	0.3696	0.31	0.5423
Contrast change	0.4305	0.4239	0.23	1.121
All Images	0.6819	0.7046	0.1112	0.9824

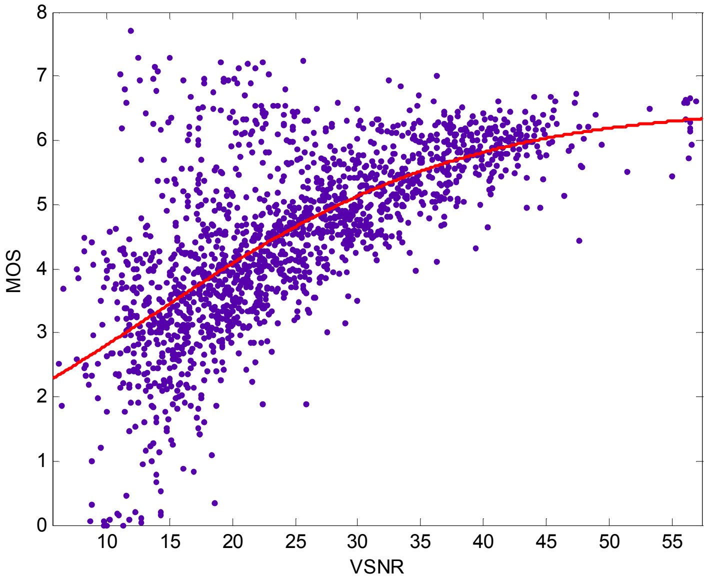


Figure 4.3.8 : Scatter plot of MOS versus VSNR on Tampere Image Database (b1=6.54, b2=0.07 ,b3=13.61)

Results in Table 4.3.8 show the performance of VSNR metric on TID2008 for each distortion type. As it can be observed from this table, VSNR has fair overall performance (pearson correlation factor is 0.6819). The better correlations between VSNR and subjective results are obtained in image denoising, JPEG and JPEG2000 compression distortion types. On the other hand, VSNR performs bad especially in localized distortions such as local block-wise distortion. One possible explanation of this result is the fact that VSNR metric does not take into account the spatial localization of distortion. In terms of consistency criteria, outlier ratio is between 0 and 0.36, and it can be concluded that the results are not consistent for all types of distortions. Figure 4.3.8 shows the relationship between MOS and VSNR pictorially in a scatter plot.

Table 4.3.9 : Performance of VSNR on LIVE Video Database for each distortion type

Distortion	**Pearson Corr.**	**Spearman Corr.**	**Outlier Ratio**	**RMSE**
Wireless	0.6975	0.6951	0	7.688
IP	0.7370	0.6929	0.033	6.914
H.264	0.6499	0.6405	0.025	8.577
MPEG2	0.5906	0.5874	0.025	8
All Videos	0.6883	0.6725	0.0133	8.043

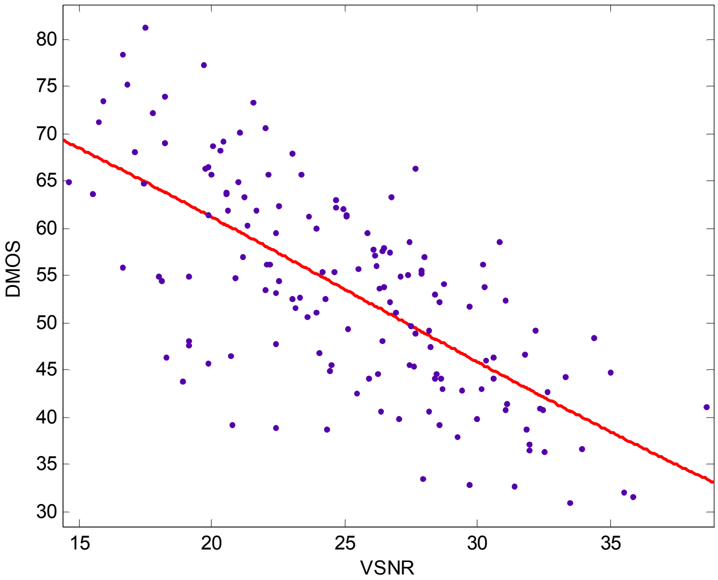


Figure 4.3.9 : Scatter plot of DMOS versus VSNR on LIVE Video Database (b1=105.3, b2=-0.29, b3=0.07)

Table 4.3.9 contains the VSNR results for all videos in LIVE Video Database classified by the distortion type. As in the TID2008, VSNR performs fair also in this database (pearson correlation factor for all videos is 0.6883). Considering VSNR model correlation with human perception for each distortion types, the best correlation is obtained in IP distortion. The worst correlation is in MPEG2 compression type of distortion. For the prediction consistency, outlier ratio for all types of distortion is between 0 and 0.033. It can be concluded that the results are consistent. Figure 4.3.9 shows the relation of subjective scores with VSNR for all the videos in LIVE Video database as a visual illustration.

4.3.5 NQM

Table 4.3.10 : Performance of NQM on Tampere Image Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Additive Gaussian noise	0.7461	0.7679	0.11	0.413
Additive noise in color component	0.7486	0.7490	0.08	0.3444
Spatially correlated noise	0.7584	0.7720	0.14	0.4142
Masked noise	0.7104	0.7067	0.14	0.4258
High frequency noise	0.9226	0.9014	0.01	0.3755
Impulse noise	0.7495	0.7616	0.1	0.3442
Quantization noise	0.8104	0.8209	0.11	0.4918
Gaussian blur	0.8789	0.8845	0.02	0.5684
Image denoising	0.9579	0.9449	0	0.4668
JPEG compression	0.9359	0.9074	0	0.6095
JPEG2000 compression	0.9409	0.9532	0.27	1.601
JPEG transmission errors	0.7320	0.7372	0.13	0.9026
JPEG2000 transmission errors	0.7339	0.7262	0.12	0.555
Non eccentricity pattern noise	0.6830	0.6799	0.17	0.7739
Local block-wise distortions	0.2185	0.2347	0.35	0.657
Mean shift (intensity shift)	0.5295	0.5245	0.2	0.4957
Contrast change	0.6817	0.6191	0.04	0.9091
All Images	0.6085	0.6243	0.1611	1.066

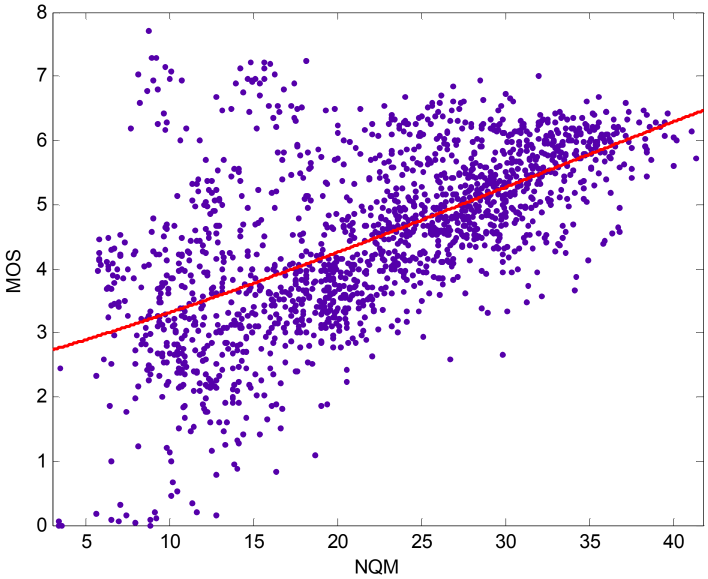


Figure 4.3.10 : Scatter plot of MOS versus NQM on Tampere Image Database
(b1=10.67, b2=0.04 ,b3=30.6)

Table 4.3.10 shows NQM results for all images in TID2008 classified by the type of distortion. As shown in table, NQM has fair overall performance (pearson correlation factor for all images is 0.6085). Considering the distortion types seperately, correlation values increases. Since NQM models the distortion on image as a combination of linear frequency distortion and additive noise injection, it performs better for such distortions. As seen in Table 4.3.10, NQM performs good in distortion types of JPEG and JPEG2000 compressions, and image denoising which can successfully be modelled. Considering the prediction consistency, for almost all types of distortion the outlier ratio is small except in local block-wise distortion. Therefore it can be stated that the results are consistent for almost all types of distortions. Figure 4.3.10 shows the relationship between MOS and NQM pictorially in a scatter plot.

Table 4.3.11 : Performance of NQM on LIVE Video Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Wireless	0.6564	0.6459	0.025	8.094
IP	0.6703	0.6667	0.0333	7.322
H.264	0.5884	0.5810	0.075	9.126
MPEG2	0.6704	0.6346	0.075	7.356
All Videos	0.6659	0.6448	0.02	8.272

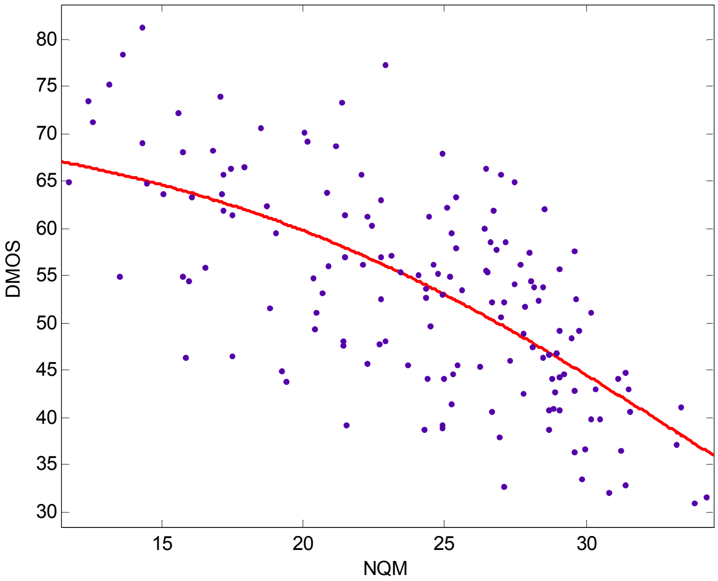


Figure 4.3.11 : Scatter plot of DMOS versus NQM on LIVE Video Database
(b1=73.16 , b2=-0.55, b3=1.89)

Table 4.3.11 presents the results of NQM metric for LIVE video database. As seen in the table, NQM performs fair in LIVE video database (pearson correlation factor is 0.6659). Regarding the distortion types, NQM is more correlated with subjective scores in MPEG2 than H.264 compression type of distortions. In addition, VIF performs better in IP distortion than in wireless distortion type. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.075, thus the results can be stated as consistent. Figure 4.3.11 shows the relationship between DMOS and NQM pictorially in a scatter plot.

4.3.6 MSSIM

Table 4.3.12 : Performance of MSSIM on Tampere Image Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Additive Gaussian noise	0.7567	0.8094	0.05	0.4361
Additive noise in color component	0.7867	0.8064	0.02	0.3217
Spatially correlated noise	0.7794	0.8195	0.01	0.4014
Masked noise	0.8072	0.8155	0	0.378
High frequency noise	0.8475	0.8685	0.01	0.5166
Impulse noise	0.6352	0.6868	0.02	0.4024
Quantization noise	0.7889	0.8537	0.02	0.521
Gaussian blur	0.9040	0.9607	0.05	0.683
Image denoising	0.9308	0.9571	0.05	1.223
JPEG compression	0.9584	0.9348	0	0.4954
JPEG2000 compression	0.9742	0.9736	0	0.4477
JPEG transmission errors	0.8607	0.8736	0	0.6803
JPEG2000 transmission errors	0.8174	0.8525	0.02	0.4708
Non eccentricity pattern noise	0.6881	0.7336	0	0.7689
Local block-wise distortions	0.7968	0.7617	0	0.4063
Mean shift (intensity shift)	0.6877	0.7374	0.02	0.4248
Contrast change	0.7688	0.6399	0	0.7944
All Images	0.8247	0.8527	0.0076	0.7636

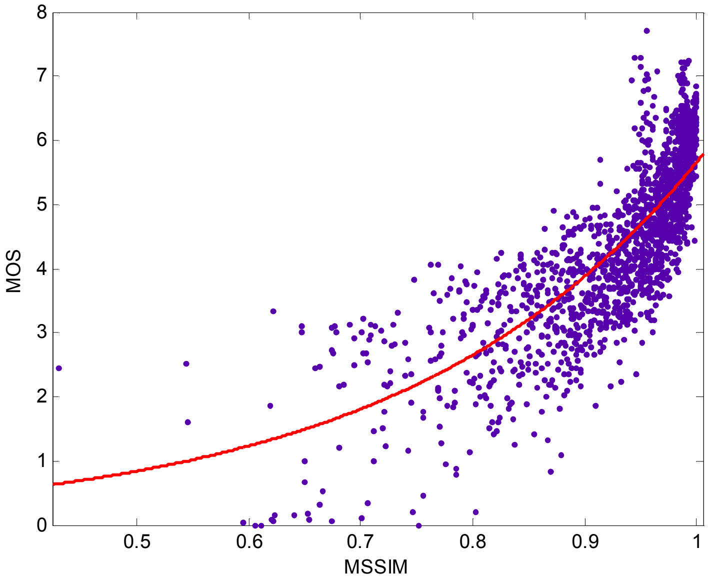


Figure 4.3.12 : Scatter plot of MOS versus MSSIM on Tampere Image Database (b1=327.7, b2=0.29, b3=14.73)

Results in Table 4.3.12 show the performance of MSSIM metric on TID2008 for each distortion type. As seen in the table, MSSIM has good overall performance (pearson correlation factor is 0.8247). The better correlations between MSSIM and HVS perception are obtained in JPEG and JPEG2000 compression distortion types. Since JPEG and JPEG2000 encode fine-scale details of image to a much higher degree than coarse-scale structures, HVS perceives better when evaluating at larger scales. This explains the better correlation after incorporating multiple scales. As seen in the table, the results are consistent by looking at the outlier ratio values which are between 0 and 0.05. Figure 4.3.12 shows the relationship between MOS and MSSIM pictorially in a scatter plot.

Table 4.3.13 : Performance of MSSIM on LIVE Video Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Wireless	0.7103	0.7291	0.025	7.552
IP	0.7233	0.6418	0	6.813
H.264	0.7321	0.7259	0.025	7.688
MPEG2	0.6781	0.6705	0	7.286
All Videos	0.7379	0.7322	0.0133	7.483

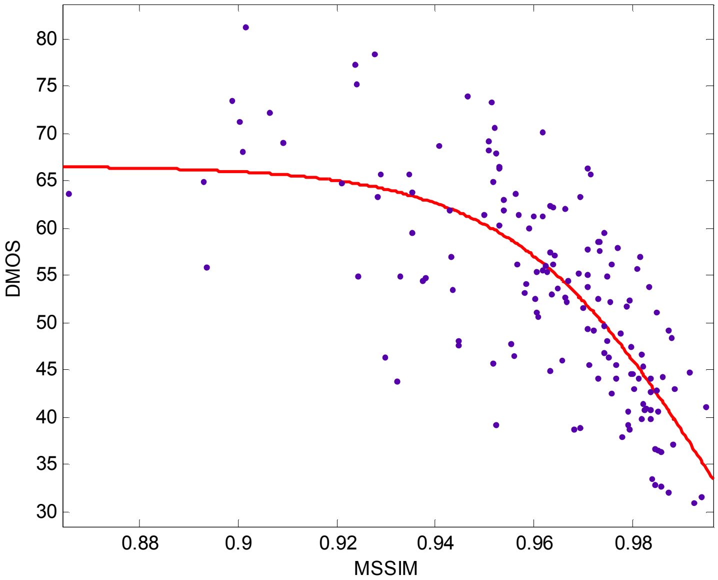


Figure 4.3.13 : Scatter plot of DMOS versus MSSIM on LIVE Video Database (b1=66.53, b2=-1.17, b3=1.46)

Table 4.3.13 presents the results of MSSIM metric for LIVE video database. As seen in the table, MSSIM performs good in LIVE video database (pearson correlation factor is 0.7379). Regarding the distortion types, MSSIM is more correlated with subjective scores in H.264 than MPEG2 compression type of distortions. In addition, SSIM performs almost the same in IP and wireless distortion types. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.025, thus the results can be stated as consistent. Figure 4.3.13 shows the relationship between DMOS and MSSIM pictorially in a scatter plot.

4.3.7 VSSIM

Table 4.3.14 : Performance of VSSIM on LIVE Video Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Wireless	0.5497	0.5502	0.05	10.73
IP	0.5712	0.4847	0.1	8.098
H.264	0.6918	0.6827	0.025	8.149
MPEG2	0.6067	0.5922	0	7.88
All Videos	0.5661	0.5485	0.046	9.14

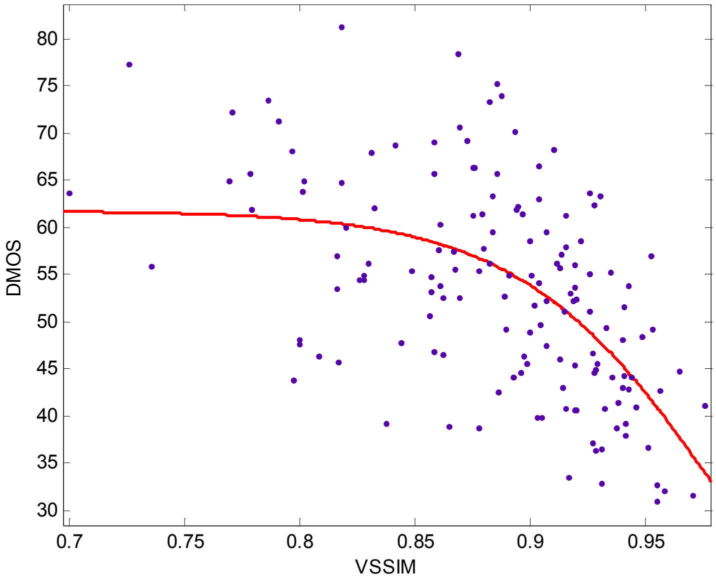


Figure 4.3.14 : Scatter plot of DMOS versus VSSIM on LIVE Video Database (b1=61.74, b2=-1.20, b3=1.86)

Table 4.3.14 shows the results of VSSIM metric for LIVE video database. As seen in the table, it performs poor in LIVE video database (pearson correlation factor is 0.5661). Regarding the distortion types, VSSIM performs better in H.264 than MPEG2 compression type of distortions. In addition, VSSIM is more correlated with HVS in IP type of distortions than in wireless. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.1, thus the results can be stated as consistent. Figure 4.3.14 shows the relationship between DMOS and VSSIM pictorially in a scatter plot.

4.3.8 VQM

Table 4.3.15 : Performance of VQM on LIVE Video Database for each distortion type

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
Wireless	0.8194	0.8009	0	6.149
IP	0.7086	0.7196	0	6.961
H.264	0.7392	0.7427	0.025	7.601
MPEG2	0.7349	0.7936	0.3	6.426
All Videos	0.7665	0.7553	0.0066	7.121

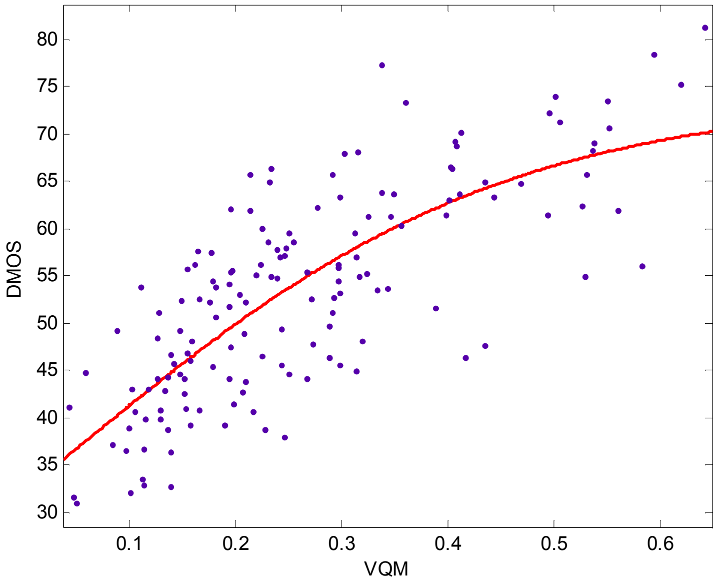


Figure 4.3.15 : Scatter plot of DMOS versus VQM on LIVE Video Database
(b1=73.91, b2=4.953, b3=0.05343)

Table 4.3.15 presents the results of VQM metric-H, The General Model, for LIVE video database. As seen in the table, VQM performs good in LIVE video database (pearson correlation factor is 0.7665). Regarding the distortion types, VQM is more correlated with subjective scores in Wireless distortion than in IP type of distortions. In addition, SSIM performs almost the same in H.264 and MPEG2 compression distortion types. Considering the prediction consistency, for all types of distortion the outlier ratio is between 0 and 0.3, thus the results can be stated as consistent. Figure 4.3.15 shows the relationship between DMOS and VQM pictorially in a scatter plot.

4.3.9 Discussion of Results

Table 4.3.16 : Performance comparison of objective quality metrics on Tampere Image Database

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
PSNR	0.5274	0.5531	0.17	1.141
SSIM	0.6449	0.6459	0.0188	1.028
VIF	0.7982	0.7496	0.0159	0.8092
VSNR	0.6819	0.7046	0.1112	0.9824
NQM	0.6085	0.6243	0.1611	1.066
MSSIM	0.8247	0.8527	0.0076	0.7636

Table 4.3.16 presents the performance of objective quality metrics on TID2008 for all images in terms of prediction accuracy, monotonicity and consistency. Comparison of objective model correlations with respect to the HVS is given as

$$PSNR < NQM < SSIM < VSNR < VIF < MSSIM$$

The results show that PSNR is the worst predictor of visual fidelity that it is not very well matched to perceived visual quality by human observer. All other objective quality assessment algorithms improve upon PSNR. Among all six algorithms the best performance is obtained in MSSIM. By this result, the notion that using multi-scale methodology can improve performance of the assessing algorithm is validated. In terms of prediction consistency, PSNR has the greater outlier ratio than all other objective quality metrics that the results are not so consistent. MSSIM has the smaller outlier ratio among all the metrics. Figure 4.3.16 shows the scatter plots of each objective quality metric versus MOS values.

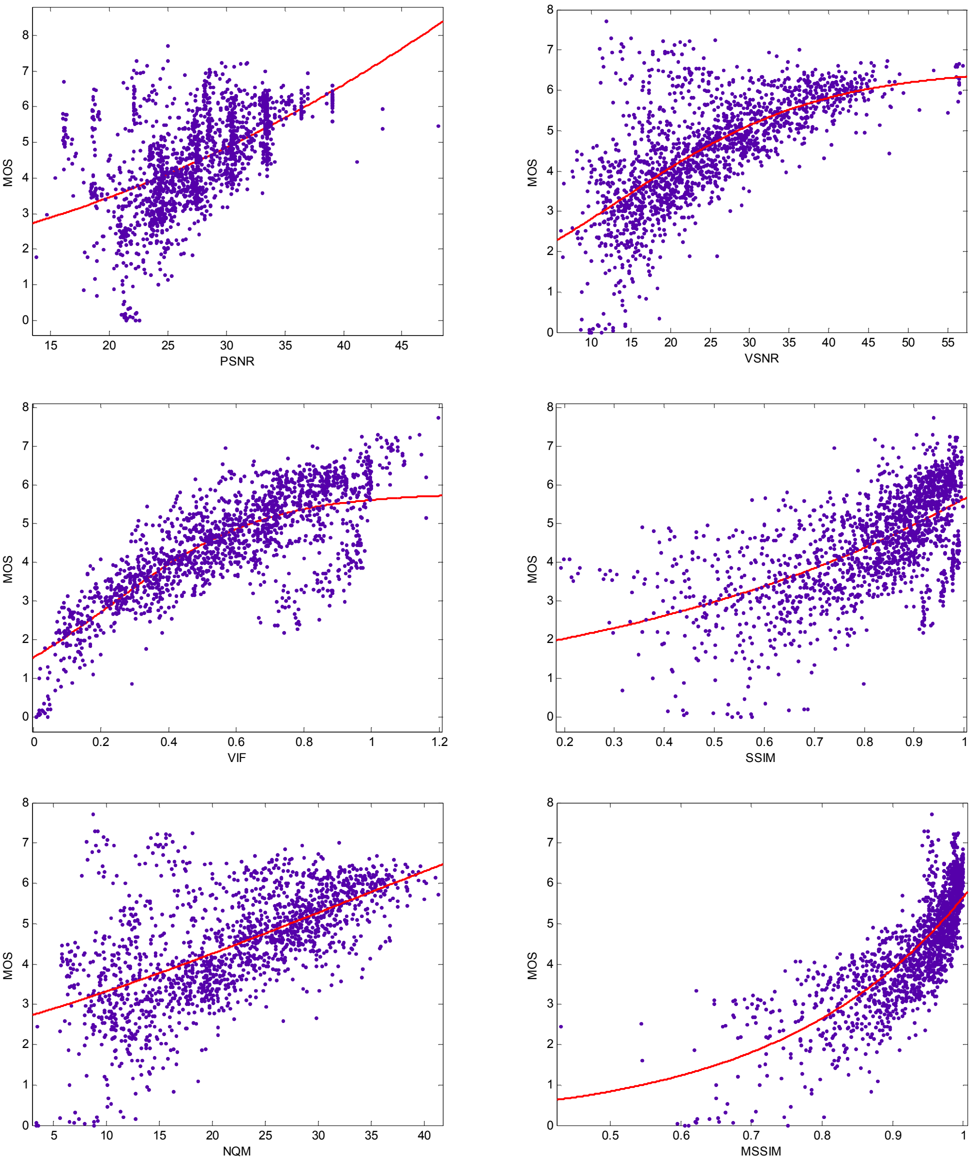


Figure 4.3.16 : Scatter plots of MOS versus objective quality metrics on Tampere Image Database

Table 4.3.17 shows the performance of objective quality metrics on LIVE video database for all videos. All the metrics except VSSIM and VQM were computed as the average of the frame level quality scores. VSSIM and VQM model the visual motion perception. Model comparison in terms of correlation to subjective results is given as

$$PSNR \approx SSIM < VIF < VSSIM < NQM < VSNR < MSSIM < VQM$$

Table 4.3.17 : Performance comparison of objective quality metrics on LIVE Video Database

Distortion	Pearson Corr.	Spearman Corr.	Outlier Ratio	RMSE
PSNR	0.5433	0.5233	0.0666	9.433
SSIM	0.5413	0.5251	0.0466	9.324
VIF	0.5541	0.5541	0.02	9.231
VSNR	0.6883	0.6725	0.0133	8.043
NQM	0.6659	0.6448	0.02	8.272
MSSIM	0.7379	0.7322	0.0133	7.483
VSSIM	0.5661	0.5485	0.046	9.14
VQM	0.7665	0.7553	0.0066	7.121

In assessing the video quality PSNR and SSIM have almost the same correlation which is poor. The best correlation performance amongst the ones in the study is achieved by VQM. Better performance of VSSIM upon SSIM illustrates the success of the visual motion perception inclusion to the algorithm. Scatter plots of all the quality metrics with respect to subjective scores are shown in Figure 4.3.17.

Performance of objective quality assessment algorithms with respect to computational efficiency is measured by taking the average time elapsed for the calculation of a frame of a video in LIVE database. Using C++ implementation for VSSIM and MATLAB for all others on 2.20 GHz Intel Core2 Duo machine, average time elapsed per frame for each metric is given in Table 4.3.18.

Table 4.3.18: Computational performance information of the quality metrics

Distortion	Time (sec)
PSNR	0.21
SSIM	0.62
VIF	7.22
VSNR	1.38
NQM	3.12
MSSIM	1.10
VQM	1.52

As seen in Table 4.3.18 the computational complexity of visual quality metrics can be given as

$$PSNR < SSIM < MSSIM < VSNR < VQM < NQM < VIF$$

VSSIM is excluded from the above comparison since it is implemented in different environment from all others. Although PSNR performs poor to match the perceived visual quality in both image and video databases, it is the simplest metric to calculate. Among all algorithms, VIF is the most complex one. VQM which is the best performing visual quality algorithm in video database has a permissible computational complexity.

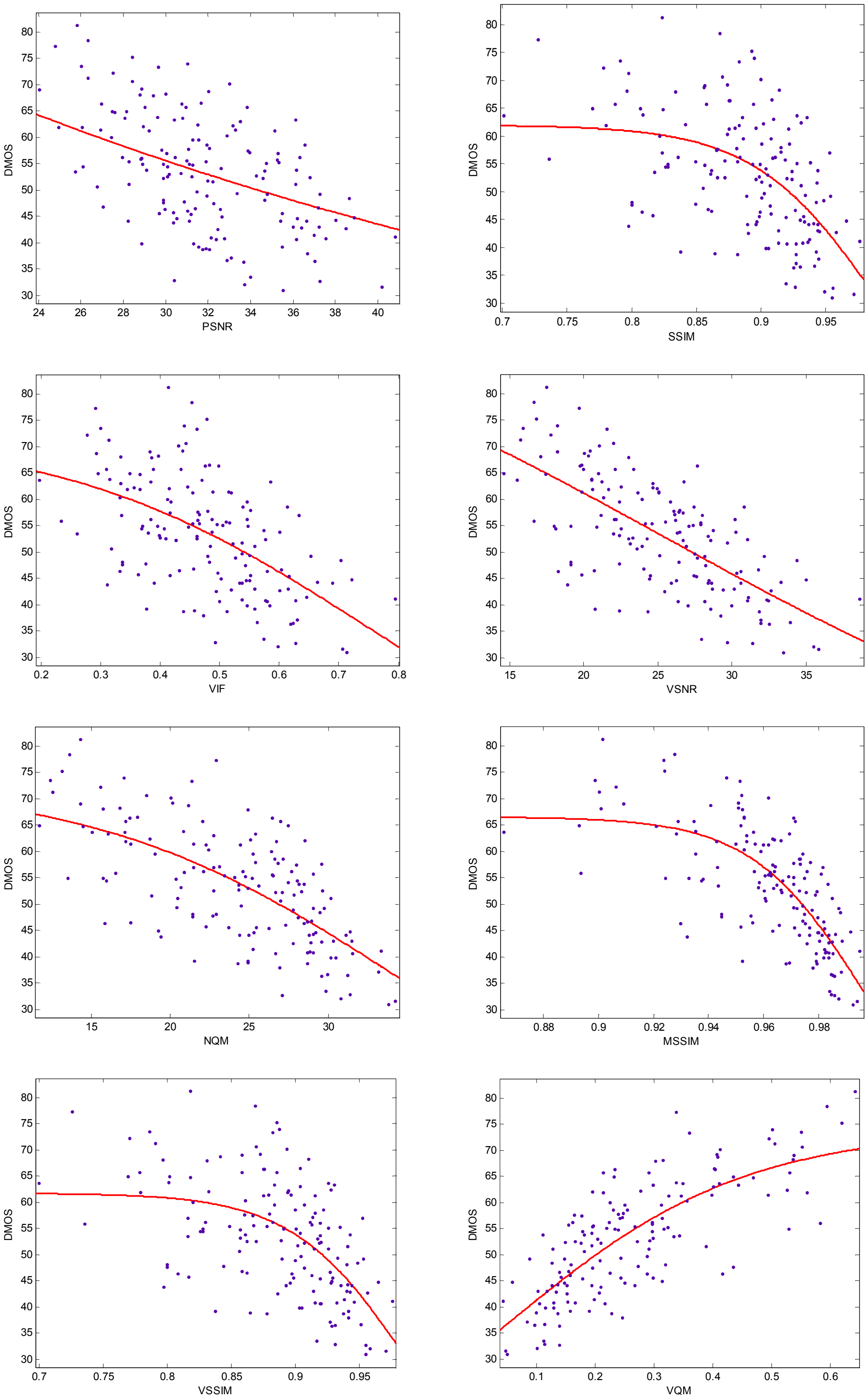


Figure 4.3.17 : Scatter plots of DMOS versus objective quality metrics on LIVE Video Database

CHAPTER 5

CONCLUSION

Throughout the study, the objective visual quality metrics that are widely accepted in literature are evaluated on a wide range of distortion types and comparison of overall performances in terms of prediction accuracy, monotonicity, consistency and computational complexity is given.

In this thesis, eight widely accepted and publicly-available objective visual quality methods are studied. Among all six metrics, VSSIM and VQM are specialized for video quality. An image and a video database which include wide variety of distortion types are downloaded and used in the evaluation of these methods. The output of the models to the databases are fitted by a logistic function to account for non-linearities in the models. Then, in order to provide quantitative performance evaluation of the objective quality metrics, methods provided in VQEG Phase II Final Report [25] are used.

The results that we obtained show that although PSNR is computationaly simple it performs very poorly against human perception when all types of distortion are taken into account at the same time. However, since PSNR assumes that the distortion is caused by additive signal-independent noise it can be useful to quantify the visual quality in such distortions.

Among all the objective quality metrics involved in the study, VQM is the best performing one in video quality evaluation. This result shows the importance of incorporating the wide range of perception based spatial-temporal features to an objective quality metric model. Taking into account also the computational complexity, VQM is an applicable metric to be employed for real world applications.

The results obtained in this study correlates well with the studies conducted before to compare performance of different visual quality metrics [19], [20], [22], [23]. A distinguishing feature of this study is the evaluation of each objective quality metrics for a wide variety of distortion types separately.

In order to evaluate the visual quality in specific applications in which the original video is partially or not available, RR and NR video quality metrics should be implemented. For future work, these type of objective metrics should be evaluated. In addition, as image and video applications continue to evolve, the distortion types that should be taken into account while designing objective metrics will increase. For example, 3D imaging has recently been widely studied. There are so many 3D applications ranging from entertainment to medical. Image quality will show the performance of these 3D systems and research on quality assessment becomes an important task. Since depth perception and stereoscopic distortions are not taken into account in 2D visual quality metrics, new objective assessment methods should be developed.

REFERENCES

[1] Jae Jeong Hwang, Hong Ren Wu and K.R. Rao "Human Visual Systems", in H.R. Wu and K.R. Rao, eds. "Digital video image quality and perceptual coding", ch. 1.6, CRC press, 2006.

[2] E. Montag and M. Fairchild "Fundamentals of Human Vision and Vision Modelling", in H.R. Wu and K.R. Rao, eds. "Digital video image quality and perceptual coding", ch. 2, CRC press, 2006.

[3] S. Winkler, Digital Video Quality – Vision Models and Metrics, John Wiley & Sons, 2005.

[4] Atanas Boev, Maija Poikela, Atanas Gotchev, and Anil Aksay, "Modelling of stereoscopic HVS", Mobile3DTV Technical report, Available: http://mobile3dtv.eu/results/, last visited on September 2011.

[5] W. Schreiber, "Fundamentals of Electronic Imaging Systems", New York, Springel, 1993.

[6] E. Peli, "Contrast in complex images," J. Opt. Soc. Amer. A, vol. 7, Oct. 1990, pp. 2032–2039.

[7] E. Peli, "In search of a contrast metric: Matching the perceived contrast of Gabor patches at different phases and bandwidths". Vision Research, 37 (23), 1997.

[8] S. Winkler and P. Vandergheynst, "Computing isotropic local contrast from oriented pyramid decompositions", in Proc. of ICIP, 4:420-424, Kobe, Japan, Oct, 1999.

[9] S. Winkler, "Perceptual video quality metrics – a review," in H.R. Wu and K.R. Rao, eds. "Digital video image quality and perceptual coding", ch. 5, CRC press, 2006.

[10] 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, Apr. 2004, pp. 600-612.

[11] 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.

[12] M. J. Wainwright, E. P. Simoncelli, and A. S. Wilsky, “Random cascades on wavelet trees and their use in analyzing and modeling natural images,” Appl. Comput. Harmon. Anal., vol. 11, 2001, pp. 89–123.

[13] 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, no. 9, 2007, pp. 2284-2298.

[14] P. J. Bex and W. Makous, “Spatial frequency, phase, and the contrast of natural images,” J. Opt. Soc. Amer. A, vol. 19, 2002, pp. 1096–1106.

[15] Damera-Venkata N., Kite T., Geisler W., Evans B. and Bovik A. "Image Quality Assessment Based on a Degradation Model", IEEE Trans. on Image Processing, vol. 9, 2000, pp. 636-650.

[16] Z. Wang and E. P. Simoncelli and A. Bovik, “Multi-scale Structural Similarity for Image Quality Assessment”, in Proc. of 37th IEEE Asilomar Conference on Signals, Systems and Computers, Nov. 2003.

[17] ITU-R BT. 500-9, “Methodology for the subjective assessment of the quality of television pictures”, 1998.

[18] ITU-T P.910, “Subjective video quality assessment methods for multimedia applications”, 1999.

[19] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, F. Battisti "Color Image Database for Evaluation of Image Quality Metrics", Proceedings of International Workshop on Multimedia Signal Processing, Australia, Oct. 2008, pp. 403-408.

[20] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, "TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics", Advances of Modern Radioelectronics, vol. 10, pp. 30-45, 2009.

[21] Tampere Image Database 2008 v.1.0. Available: http://www.ponomarenko.info /tid2008.htm, last visited on September 2011.

[22] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video", IEEE Transactions on Image Processing, vol.19, no.6, pp.1427-1441, June 2010.

[23] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "A Subjective Study to Evaluate Video Quality Assessment Algorithms", SPIE Proceedings Human Vision and Electronic Imaging, Jan. 2010.

[24] LIVE Video Quality Database. Available: http://live.ece.utexas.edu/research/ quality/live_video.html, last visited on September 2011.

[25] Final report from the Video Quality Expetrs Group on the validation of objective models of video quality assessment, Phase II, VQEG, August 2003.

[26] 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, Feb. 2004, pp. 121-132.

[27] B. Girod, “What’s wrong with mean-squared error,” in Digital Images and Human Vision, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 207–220.

[28] Final report from the Video Quality Expetrs Group on the validation of objective models of video quality assessment, VQEG, March 2000 .

[29] Curve Fitting Toolbox for MATLAB, Available: http://www.mathworks.com/products/curvefitting/index.html, last visited on September 2011.

[30] Metrix Mux Visual Quality Assessment Package, Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/, last visited on September 2011.

[31] M. Pinson and S. Wolf. "A New Standardized Method for Objectively Measuring Video Quality," IEEE Transactions on Broadcasting, vol. 50, no. 3, September, 2004, pp. 312-322.

APPENDIX-A

SOFTWARE TOOLS

A.1 Metrix MUX

MeTriX MuX is a Matlab package that implements wrapper code for a variety of visual quality assessment algorithms. The algorithms currently supported by the package are listed below:

- MSE
- PSNR
- SSIM
- MSSIM
- VSNR
- VIF
- VIFP
- UQI
- IFC
- NQM

The package is publicly-available to download [30] and all the instructions and information are included in it.