AN ABSTRACT OF THE THESIS OF

Simone Leorin for the degree of Master of Science in

Electrical and Computer Engineering presented on May 27, 2005.

Title: Quality Assessment Strategies for Multi-Camera Panorama Video

Abstract approved: _____

Luca Lucchese

New video-conference devices based on omnidirectional multi-camera systems have been emerging in the last few years. These devices require innovative and automated video quality assessment in the earlier stages of their design in order to guarantee competitive product development and quality monitoring. Current quality assessment techniques are not adequate since they are mostly tailored to single video cameras. Even if these techniques are capable of assessing the quality of each video stream separately, the overall quality of a panorama video stream generated with the outputs of multiple cameras stitched together presents strong deviations from the results of subjective quality tests. In this thesis, we present new strategies for assessing the quality of panorama video streams with specific emphasis on the following problems: noticeable calibration differences between adjacent cameras, concentration of motion in limited regions of the panoramic scene, combined vignetting problems, non-uniformity of the surfaces in the seam region between two adjacent cameras. Combining different features from high and low level vision, we evaluate the proposed perceptual quality metric using a set of customized test sequences and verify its correlation with subjective quality tests.

Quality Assessment Strategies for Multi-Camera Panorama Video

by

Simone Leorin

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented May 27, 2005
Commencement June 2006

Master of Science thesis of Simone Leorin presented on May 27, 2005

APPROVED:

Major Professor, representing Electrical and Computer Engineering

---

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Simone Leorin, Author

ACKNOWLEDGMENTS

I am deeply indebted to my advisor, Professor Luca Lucchese, for his guidance and encouragement and, most importantly, his friendship. He was the one directed my interest toward notion of commitment and who made sure throughout the whole process that I had all the resources I needed to carry out my research. Furthermore, his impressive skills in the kitchen have often led us to share our research findings during unforgettable dinners.

I also want to express my gratitude to the other members of my Committee: Professor Mario E. Magaña, Professor Thinh Nguyen, and Professor Jimmy Yang. Their advice and patience is greatly appreciated.

I am grateful to Dr. Ross Cutler at Microsoft Corp. who enabled me to gain the much-needed industrial experience working in a professional software development organization. A special acknowledgment is due to Harsh Nanda, who made me realized that successful software development does require much more than just technical knowledge.

I wish to express my most sincere gratitude to Sundar, Nelson, Guddi, Krishna, Tamal, Soumit, Saswata, Kumar, friends and colleagues of mine. The claims made in this thesis could not have been empirically explored or validated without their help. Our poker nights have also been invaluable to enjoy and share good time together.

I also wish to express my gratitude to Robert and David, friends of mine following a similar path at OSU with whom I had the opportunity to share the joys and the frustrations of doing research. Their friendship has been fundamental in many occasions.

I specially wish to thank my family for their continuous support throughout my life in all my aspirations, whether these have made sense or not.

Finally, a very special thank to my wife Simona, whose invisible contribution to the completion of this thesis is written large in long periods of distance and sacrifices. She has always been my source of energy and a constant presence.

Simone Leorin

Redmond, May 20th 2005

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

QUALITY ASSESSMENT STRATEGIES FOR MULTI-CAMERA PANORAMA VIDEO

1. CHAPTER ONE

1.1. Introduction

With the availability of optimized audio/video streams and improved network resources, new devices have recently started entering the market, promising unprecedented videoconferencing experiences. By combining the outputs of multiple video sources (see Fig. 1.1), a new video format is created and its quality needs to be evaluated by using novel strategies. Providing the best trade-off between technical requirements and appealing video quality is still the main concern in the design process of such devices. Therefore, as the end users are getting increasingly accustomed to video technologies and, consequently, more demanding, the perceived quality of digital video is a very important factor in discriminating between successful video streaming technologies.

This research focused on the adjustment and enhancement of current quality assessment techniques to appropriately address specific problems experienced during the quality test of a new videconferencing device currently being developed at Microsoft Corporation. Classical quality assessment approaches have revealed their weaknesses because they are mostly designed to evaluate the quality of standard video formats. These techniques have demonstrated the capability of adequately assessing the quality of single video streams. However, the overall quality assessed for a panorama video stream, generated by stitching together the

FIGURE 1.1. An example of uncalibrated panoramic video stream generated by an omnidirectional device.

outputs of multiple cameras, presents low correlation with the results of subjective quality tests. Problems such as noticeable calibration differences between adjacent cameras, concentration of motion in limited regions of the panoramic scene, combined vignetting, non-uniformity of the surfaces in the seam regions, are not appropriately handled by existent metrics. In this thesis, we present new strategies for assessing the quality of panorama video with specific attention to those peculiar problems that have challenged standard approaches.

A short description of this work was given in this introductory paragraph. Section 1.2 reviews the purposes of the quality metrics and basic terminology of video quality assessment. Section 1.3 presents an overview of the Thesis.

1.2. Video Quality Assessment

The goal of objective video quality assessment is to design quality metrics that can predict perceived video quality automatically. In general, the purpose of an image/video quality metric is threefold [1]:

- *Monitoring:* A quality metric can be used to monitor image quality for quality control systems. For example, an image and video acquisition system can use the quality metric as an automatic control to adjust itself to obtain

the best image/video quality data. A network video server can examine the quality of the digital video transmitted on the network and regulate the video streaming accordingly.

- *Benchmarking:* A quality metric can be employed to benchmark image and video processing systems and algorithms. If multiple video processing systems are available for a specific task, then a quality metric can help determine a ranking and provide the best quality results.

- *Optimizing:* In the early stages of design, where optimization is one of the key issues, a quality metric can be embedded in the image/video system to optimize the algorithms and the parameter settings. For instance, in the multicamera system under analysis, a quality metric can help optimal design of the post-filtering and calibration algorithms at the stitcher and decoder level.

It is important to point out that there are two notions usually associated with the concept of quality. First, quality implies a comparison between two or more quantities or objects. This comparison may be direct (A is better than B) or indirect (A is good). Second, given the possibility of something being better or worse, quality must be quantified using an open-ended scale. These two simple concepts have a considerable impact on how video quality is measured. In fact, the fundamental distinction that can be made on objective video quality metrics is based on the full or partial availability of the original (distortion-free) video signal, with which the processed signal is to be compared:

- Full-Reference (FR): A complete reference image/video is assumed to be known; these objective measures of impairment are more appropriately

termed fidelity measures [2] and are based on the differences between source and processed signal;

- Reduced-Reference (RR): The reference image/video is only partially available in the form of a set of extracted features made available as side information to help the evaluation process. This approach usually adopts an ancillary channel for reduced key features interchange;

- No-Reference (NR)(also referred to as "blind" quality assessment): The reference image/video is not available at all.

In most of the objective quality metrics proposed in the literature, the undistorted reference signal is supposed to be fully available. But it is worth noting that, in many practical video service applications, the reference images or video sequences are often not accessible. Therefore, it is highly desirable to develop measurement approaches that can evaluate image and video quality blindly. In the case under investigation, for instance, the impossibility to record the video remotely (after being transmitted through the network) has motivated us to develop an NR metric that assesses the quality indirectly.

As subjective tests have largely demonstrated, a human observer can straightforwardly judge the quality of a processed image or video without taking into consideration the original as a reference. On the other hand, developing objective quality assessment methods that do not require a reference, is still a quite difficult task. The user judgment maintains an unquestionable importance in the final quality evaluation. The Mean Opinion Score (MOS) quality measurement obtained from human observers has been regarded for many years as the most reliable form of quality measurement. However, the drawbacks of this approach include high cost, high time consumption, and practical inconvenience. It

is therefore important to develop objective methods able to automatically predict these subjective evaluations and establish a consistent correlation with numerical data.

This work will focus on the development of an adjusted NR metric but, for the sake of clarity, some notions helpful in the three domains (FR,NR,RR) will be presented in this and in the following chapter.

The most widely used FR objective image and video distortion/quality metrics are the Mean Squared Error (MSE) and the Peak Signal-to-Noise Ratio (PSNR), which are defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{1.1}$$

and

$$PSNR = 20\log_{10}\frac{V_{peak}}{MSE} \tag{1.2}$$

where N is the number of pixels in the image or video signal, and x_i and y_i are the $i-th$ pixels in the original and the distorted signals, respectively. V_{peak} is the dynamic range of the pixel values (255 for an 8 bits/pixel monotonic signal). Based on the sum of all differences between the video and the reference one, MSE and PSNR are widely used because they are simple to calculate, have clear physical meanings, and are mathematically easy to deal with (MSE is differentiable, for example). However, they have been widely criticized as well for not correlating well with perceived quality measurements [1, 3–8]. These simple measures operate solely on a pixel-by-pixel basis and neglect the important influence of image content and viewing conditions on the actual visibility of artifacts. Therefore, their predictions do not agree well with perceived quality. Considerable efforts have

been made to develop objective image/video quality assessment methods based on the human visual system (HVS) characteristics. Some of the developed models are commercially available. Even though only limited success has been reported from HVS-based FR quality assessment models [6, 9], there is an increasing interest in these approaches. HVS-models may slowly replace classical schemes, in which the quality metric basically consists of an MSE or PSNR measure. The quality improvement that can be achieved using an HVS-based approach is significant and applies to a large variety of image processing applications. Although HVS models can be very complex and computationally demanding (inconvenient for DSP implementations), simplified versions can be extremely powerful. As discussed later, standard quality metrics applied to panorama video have revealed weaknesses that have been appropriately tackled by simplified HVS-based semantic models.

1.2.1. Video Quality Metrics

The quality metrics measure the typical artifacts introduced by processing and transmitting digital video. To be accurate, digital video quality measurements must be based on the perceived quality of the actual video received by the users rather than on the measured quality of traditional video test signals (e.g., color bar used in analog video testing). In fact, digital video systems adapt and change their behavior depending upon the dynamic features of both video content (e.g., spatial information, motion) and digital transmission system (e.g., bit rate, dropped packets, latency). Therefore, different types of impairments than those generated by analog video devices have to be taken into consideration.

Widely used metrics for video quality assessment are the ANSI video quality metrics and the perceptual metrics. ANSI metrics rely on features and parameters suggested by the American National Standards Institute (ANSI) [10–12] and include common fidelity metrics (such as the MSE and the PSNR) and spatiotemporal metrics (such as the Edge Energy Difference and the Motion Energy Difference). They represent arithmetic measures of the distance between processed and reference video. Although very popular in the image and video processing community, ANSI metrics do not take into account human perception. On the other hand, perceptual metrics measure video artifacts as perceived by the viewer. These metrics aim to measure impairments (such as jerkiness, blockiness, blur, noise, colorfulness) in a way that is correlated with human perception of those impairments. An additional characteristic of some perceptual metrics is that they can be calculated without the reference video. Being the reference video not available in the scenarios under investigation, perceptual metrics have thus been the primary methods for the development of the adjusted metric proposed in this work. It is worth noting that modified ANSI metrics have been successfully used on subsequent 'static' frames. Spatiotemporal constant regions of the video have been analyzed with these techniques to calculate noise, to identify spatio/temporal local fluctuation (artifacts such as flickering), and to quantify and track motion.

Some background details on the two categories of metrics are given in the rest of this paragraph.

1.2.1.1. ANSI Video Quality Metrics

These metrics measure differences between the original (source) video and the received (destination) video. The amount of different types of distortion that has occurred in the transmission process can be calculated directly from the destination video. As previously stated, being the source video not accessible, these techniques have been used on the destination video only. Differences have been measured between subsequent frames with similar spatiotemporal characteristics, identifying noise and artifacts.

The parameters suggested by ANSI are divided in three main categories: parameters based on scalar features, parameters based on vector features, and parameters based on matrix features. A brief description of the basic concepts is given below and some examples are given in Appendix A.

Parameters based on scalar features

Scalar, or one-dimensional, features produce one value per video frame. ANSI has defined a set of scalar quality parameters that can be extracted from the source and destination videos. Two useful concepts that are used in the computation of several scalar quality parameters are the *Spatial Information* (SI), and the *Temporal Information* (TI).

Spatial Information $SI(i, j, n)$

Spatial gradients, or edges, play an important role in image quality and can be enhanced using different edge-enhancing filters. Sobel filters are suggested in the ANSI document [12]. In this case, the edge-enhanced images are obtained by linearly convolving each video frame with the kernels in Eq. (1.3):

$$H_v = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}, \; H_h = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}. \tag{1.3}$$

The results of these convolutions are

$$SI_v(i,j,n) = Y(i,j,n) \otimes H_v \tag{1.4}$$

and

$$SI_h(i,j,n) = Y(i,j,n) \otimes H_h, \tag{1.5}$$

where $Y(i,j,n)$ denotes the luminance component at pixel location (i,j) of the $n{-}th$ frame, and $SI_v(i,j,n)$ and $SI_h(i,j,n)$ are called vertical and horizontal spatial information of the video frame, respectively.

Parameters values based on spatial information can be interpreted as indication of added or lost edges in the destination scene compared to the source scene. Added edges result from impairments such as tiling, error blocks, or noise. On the other hand, lost edges may result from blurring.

We define the magnitude (radius) and phase of the spatial information respectively as

$$SI_r(i,j,n) = \sqrt{SI_v^2(i,j,n) + SI_h^2(i,j,n)} \tag{1.6}$$

and

$$SI_\Theta(i,j,n) = \arctan\Big(\frac{SI_v(i,j,n)}{SI_h(i,j,n)}\Big). \tag{1.7}$$

Some statistical properties about the spatial information can be obtained by computing the mean, variance, standard deviation, and RMS of $SI(i,j,n)$:

$$SI_{mean}(n) = \frac{1}{P} \sum_i \sum_j SI_r(i,j,n) \; , \tag{1.8}$$

$$SI_{var}(n) = \frac{1}{P}\sum_i \sum_j (SI_r(i,j,n) - SI_{mean}(n))^2 \ , \tag{1.9}$$

$$SI_{stdv}(n) = \sqrt{SI_{var}(n)}\,, \tag{1.10}$$

$$SI_{rms}(n) = \sqrt{SI_{var}(n) + SI^2_{mean}(n)}\,, \tag{1.11}$$

where P is the total number of pixels.

Temporal information $TI(i,j,n)$

Temporal information $TI(i,j,n)$ as defined in Eq. (1.12), describes the difference (movements) between two temporally adjacent frames $Y(i,j,n-1)$ and $Y(i,j,n)$:

$$TI(i,j,n) = Y(i,j,n) - Y(i,j,n-1). \tag{1.12}$$

Quality parameters based on temporal information can be interpreted as indication of added or lost motion in the destination scene compared to the source scene. Added motion results from impairments such as jerkiness, error blocks, and noise. Frame repetition clearly leads to lost motion. In our case, temporal information is extracted from subsequent frames, assumed to be constant, in order to detect unwanted added motion.

As with the spatial information feature, we can get valuable information by computing the mean, variance, standard deviation, and RMS of $TI(i,j,n)$:

$$TI_{mean}(n) = \frac{1}{P}\sum_i \sum_j TI_r(i,j,n)\,, \tag{1.13}$$

$$TI_{var}(n) = \frac{1}{P}\sum_i \sum_j (TI_r(i,j,n) - TI_{mean}(n))^2\,, \tag{1.14}$$

$$TI_{stdv}(n) = \sqrt{TI_{var}(n)}\,, \tag{1.15}$$

$$TI_{rms}(n) = \sqrt{TI_{var}(n) + TI^2_{mean}(n)}\,, \tag{1.16}$$

where P is the total number of pixels.

Examples of scalar quality parameters based on $TI(i,j,n)$ and $SI(i,j,n)$ are given in Appendix A.

Parameters based on vector features

Let us define a square subregion $R(i,j,n)$ of $Y(i,j,n)$ comprised of $N \times N$ pixels. A vector feature can now be derived from the magnitude of the Fourier transform of $R(i,j,n)$:

$$F(k,l,n) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} R(i,j,n) e^{-j\frac{2\pi}{N}(ki+lj)}\,. \tag{1.17}$$

Then, the average of the N_r magnitudes of spatial frequency components that lie within a certain discrete radius r is computed as

$$f(r,n) = \frac{1}{N_r} \sum_{i,j} |F(k,l,n)|, \tag{1.18}$$

and a vector-valued feature is defined as

$$f(n) = \begin{pmatrix} f(0,n) \\ f(0,1) \\ \vdots \\ f(\frac{N}{2}-1,n) \end{pmatrix}. \tag{1.19}$$

The vector-valued feature described in Eq. (1.19) is computed for both the source and the destination scene. Two parameters that can be computed from these

vectors are, for example, the Maximum (over time) of the Summed Positive Error Ratio (SPER) and the Minimum (over time) of the Summed Negative Error Ratio(SNER) defined as

$$SPER = \max_{n} \left(\sum_{i=il}^{iu} \left[\frac{f_S(i,n) - f_D(i,n)}{f_S(i,n)} \right]_{PositivePart} \right) \tag{1.20}$$

and

$$SNER = \min_{n} \left(\sum_{i=il}^{iu} \left[\frac{f_S(i,n) - f_D(i,n)}{f_S(i,n)} \right]_{NegativePart} \right). \tag{1.21}$$

where $f_S(i,n)$ and $f_D(i,n)$ are the vector-valued features computed for the source and the destination video, respectively.

Parameters based on matrix features

One common parameter used for signal quality is the signal-to-noise ratio. Because of the two-dimensional (matrix) nature of a digital picture, the SNR of an image can be considered as a matrix-based quality parameter. The peak signal-to-noise ratio in dB of a digital image is defined as

$$PSNR(n)_{dB} = 20 \log 10 \left(\frac{V_{peak}}{\sqrt{\frac{1}{N_{col} N_{row}} \sum_{i=0}^{N_{col}-1} \sum_{j=0}^{N_{row}-1} [Y_S(i,j,n) - Y_D(i,j,n)]^2}} \right), \tag{1.22}$$

where $Y_S(i,j,n)$ and $Y_D(i,j,n)$ are the luminance components of the source and the destination video, respectively. The word *peak* refers to the maximum value of a pixel. In the case of eight bits per pixel, $Vpeak$ in Eq. (1.22) is $2^8 - 1 = 255$.

1.2.1.2. Perceptual Metrics

Perceptual quality metrics quantify the artifacts present in the video as perceived by human viewers. These well-known artifacts can be easily recognized

even by non-experts. The most common artifacts are briefly presented here and summarized in Table 1.1. The purpose of these metrics is to provide an automatic measure of those artifacts in order to establish a correlation with human perception. Since some of them are specifically designed to assess the quality in the absence of a reference (jerkiness, blockiness, blur, noise), these metrics have been extensively used in this work to predict the MOS and reproduce the results of human subjective tests.

Jerkiness. Jerkiness is a perceptual measure of frozen pictures or motion that does not look smooth. The primary causes of jerkiness are network congestion and/or packet loss. It can also be introduced by the encoder dropping or repeating entire frames in an effort to achieve the given bit-rate constraints. A reduced frame rate can also create the perception of jerky video. Lower levels of jerkiness can be perceived when subregions of the image appear to be moving in a jerky way.

Blockiness. Blockiness is a perceptual measure of the block structure that is common to all image compression techniques based on the Discrete Cosine Transform (DCT). The DCT is typically performed on 8×8 blocks in the frame and the coefficients in each block are quantized separately, leading to artificial horizontal and vertical borders between these blocks. Blockiness can also be caused by transmission errors, which often affect entire blocks in the video.

Blur. Blur is a perceptual measure of the loss of fine detail and the smearing of edges in the video. It is due to the attenuation of high frequencies at some stage of the recording or encoding process. It is one of the main artifacts of wavelet-based compression techniques, such as JPEG2000, where transmission errors or packet loss can also induce blur. DCT-based compression schemes are also affected by this artifact, although to a lesser extent (JPEG, MPEG). Other

important sources of blur are low-pass filtering, out-of-focus cameras, and large motion (leading to motion blur).

Noise. Noise is a perceptual measure of high-frequency distortions in the form of spurious pixels. It is more noticeable in smooth regions and around edges (edge noise). This can arise from noisy recording equipment, from the compression process (where certain types of image content introduce noise-like artifacts), and from transmission errors (especially uncorrected bit errors).

Ringing. Ringing is a perceptual measure of ripples, typically seen around high-contrast edges in otherwise smooth regions (this is referred to as the Gibbs phenomenon). Ringing artifacts are very common in wavelet-based compression schemes (e.g., JPEG2000), but they also appear to a slightly lesser extent in DCT-based compression techniques (e.g., JPEG, MPEG).

Colorfulness. The colorfulness of an image describes the intensity or saturation of colors as well as the spread and distribution of individual colors in the image. The range and saturation of colors is often impaired by compression.

Watermarking Artifacts. Digital watermarking of digital video content is becoming an increasingly important way for content producers to protect their production and distribution. It is important to minimize the perceptual impact of the embedded watermark on the content. The ideal way to do this is to use perceptual metrics that can reproduce the impact of the watermark on a human observer.

1.3. Thesis overview

The thesis is organized as follows:

- Chapter 1 presents basic concepts and definitions behind FR, NR, and RR image and video quality assessment.

- Chapter 2 reviews current literature on image/video quality assessment. A set of algorithms chosen to provide an initial quality estimation of the video is also discussed.

- Chapter 3 introduces the proposed adjusted NR metric for assessing the quality of panorama video streams, illustrating the specific drawbacks encountered using standard methods. The proposed perceptual quality metric is evaluated using a set of customized test sequences and its correlation with subjective quality tests is verified. Some preliminary results of the novel NR metric developed are included. The chapter finally addresses issues regarding the prediction performance of the quality metric.

- Chapter 4 makes concluding remarks and provides suggestions for future research directions.

TABLE 1.1. Perceptual Metrics

Metric	**Jerkiness(FR, NR, Temporal)**
Description	Frozen pictures or motion that does not look smooth.
Common Cause	Network congestion, packet loss, dropped frames reduced frame rate.
Metric	**Ghosting(FR, Temporal)**
Description	Delayed version of the picture appearing on the screen.
Common Cause	Network congestion, packet loss, dropped frames reduced frame rate.
Metric	**Blur(FR, NR, Spatial)**
Description	Loss of fine detail and edge smearing due to high-freq. attenuation.
Common Cause	Compression (wavelets, DCT), transmission error, packet loss, low-pass filtering, camera out-of-focus, high motion.
Metric	**Blockiness(FR, NR, Spatial)**
Description	Block grid structure (discontinuities at adjacent block boundaries).
Common Cause	Compression (wavelets, DCT), transmission error, packet loss, low-pass filtering, camera out-of-focus, high motion.
Metric	**Ringing(FR, Spatial)**
Description	Ripples around high-contrast edges in otherwise smooth regions.
Common Cause	Wavelet compression (e.g. JPEG2000), DCT compression.
Metric	**Colorfulness(FR, Spatial)**
Description	Loss of color.
Common Cause	Compression, different processing of luminance and color.
Metric	**Noise(FR, NR, Spatial)**
Description	High frequencies distortions (spurious pixels).
Common Cause	Noisy recording equipment, compression, motion compensation mismatch, transmission errors.

2. CHAPTER TWO

2.1. Video Quality Assessment Algorithms

A video signal can be thought of as a sum of a perfect reference signal and an error signal. We may assume that the loss of quality is directly related to the strength of the error signal. Therefore, a natural way of assessing the quality of a video is to quantify the error between the distorted signal and the reference signal, which is fully available in FR quality assessment. We have seen in the previous chapter that the simplest implementation of the concept is the MSE as given in Eq. (1.1). However, there are a number of reasons why the MSE may not correlate well with the human perception of quality:

1. Digital pixel values on which the MSE is typically computed may not exactly represent the light stimulus entering the eye.

2. The sensitivity of the HVS to the errors may be different for different types of errors and may also vary with visual context. This difference may not be captured adequately by the MSE.

3. Simple error summation, like the one implemented in the MSE formulation, may be markedly different from the way the HVS and the brain arrive at an assessment of the perceived distortion.

In the last three decades, many of the proposed image and video quality metrics have tried to improve the MSE by addressing the above issues. They have followed a paradigm error sensitivity based, which attempts to analyze and quantify the error signal in a way that simulates the characteristics of human visual error perception. Pioneering work in this area was done by Mannos and

Sakrison [43] and has been extended by other researchers over the years [18, 28, 40, 46, 48]. Without providing details about the HVS, which is beyond the scope of this thesis, it is worth mentioning that the underlying principle of visual error sensitivity-based algorithms is to predict perceptual quality by quantifying perceptible errors. This is accomplished by simulating the functional components of the HVS involved in the perceptual quality evaluation. However, the HVS is an extremely complicated system, whose understanding is currently limited. Therefore, many visual error sensitivity-based approaches, explicitly or implicitly, make a number of arguable assumptions (perfect quality of the reference signal, complete parametrization of the eye's optic, suppression of the active visual processes, etc.).

A new interesting philosophy has recently emerged [1, 3, 7] which focuses on the principle that the main function of the human visual system is to extract structural information from the viewing field, and that the HVS is highly adapted for this purpose. Since a modified version of this approach has been used also in this work, a more detailed mathematical formulation of the method is presented in Section 2.2.3. The basic idea behind the metric is that a measurement of structural distortion should be a good approximation of perceived image distortion. In other words, image degradations are considered as perceived structural information loss instead of perceived errors. Although errors and structural distortion sometimes are in agreement, in many circumstances the same amount of error may lead to significantly different structural distortion. This leads to a quite effective demonstration of the limitations of the MSE to quantify the quality of an image.

The motivating example is shown in Fig. 2.1 and Fig. 2.2 where the original "Lena" image is altered with a wide variety of distortions such as mean shift,

FIGURE 2.1. Evaluation of "Lena" images with different types of distortion [8]. Top-left: Mean shifted image, $MSE = 225$, $Q = 0.9894$; Top-right: Contrast stretched image, $MSE = 225$, $Q = 0.9372$; Bottom-left: Blurred image, $MSE = 225$, $Q = 0.3461$; Bottom-right: JPEG compressed image, $MSE = 215$, $Q = 0.2876$.

contrast stretching, blurring, heavy JPEG compression, and various types of noise (salt & pepper, additive Gaussian noise, and speckle noise).

Tuning all the distorted images to yield almost the same MSE in relation to the original image, it is interesting to notice that images with nearly identical MSE have drastically different perceptual quality. Simple subjective evaluation assessment shows that the contrast stretched and the mean shifted images provide

FIGURE 2.2. Evaluation of "Lena" images with different types of noise [8]. Top-left: Original "Lena" image, 512×512, 8 bits/pixel; Top-right: Impulsive salt-pepper noise contaminated image, $MSE = 225$, $Q = 0.6494$; Bottom-left: Additive Gaussian noise contaminated image, $MSE = 225$, $Q = 0.3891$; Bottom-right: Multiplicative speckle noise contaminated image, $MSE = 225$, $Q = 0.4408$.

very high perceptual quality, while the blurred and the JPEG compressed images have the lowest subjective scores.

This is not surprising with a good understanding of the new approach since the structural change from the original to the contrast stretched and mean shifted images is trivial, but the structural change for the blurred and JPEG compressed images is very significant.

As already pointed out, in our case the reference video is not available. This impediment to the feasibility of an FR video quality assessment has affected our approach to the problem. The RR quality assessment was not a feasible approach either, because it was not possible to implement an ancillary data channel for reduced features interchange.

At this point, given the limited success that FR quality assessment has achieved, it should come as no surprise that designing objective no-reference (NR) quality measurement algorithms is very difficult indeed. This is mainly due to the limited understanding of the HVS and the associated cognitive aspects of the brain. Only a few methods have been proposed in the literature [13–17, 22, 29, 37] for objective NR quality assessment; yet, this topic has recently attracted a great deal of attention. For example, the video quality experts group (VQEG) [37] considers the standardization of NR and RR video quality assessment methods as one of its working directions. More details on the VQEG are given in Appendix A.

Furthermore, the problem of NR quality assessment is made even more complex by the fact that many unquantifiable factors play a role in human assessment of quality, such as aesthetics, cognitive relevance, learning, visual context, etc., when the reference signal is not available for MOS evaluation. These factors introduce variability among human observers based on individual subjective impressions, which has to be handled in some way. However, we can work with the following paradigm for NR quality assessment: all images and videos are flawless unless distorted during acquisition, processing, and reproduction. Hence, the task of blind quality measurement simplifies into blindly measuring the distortion that has possibly been introduced during the stages of acquisition, processing, and re-

production. The distortion is separated from the "expected" signals by making assumptions regarding statistics of "perfect natural images." For example, natural images do not contain blocking artifacts, and therefore any presence of periodic edge discontinuity in the horizontal or vertical directions is probably a distortion introduced by block-DCT based compression techniques.

2.2. Towards a New Approach

In the case under investigation, in addition to common artifacts present in digital video streams such as blur, blockiness, noise, and ringing [18], new impairments are introduced by the specific technology employed. The former are primarily due to compression [19] and network conditions as well as to non-trivial interactions between the spatial/temporal characteristics of the video sequence and the type of codec used [20]. The latter, as we will see in the next chapter, are mainly the result of combining the video signals of different sources into one single multimedia stream. The video streams from five different cameras are stitched together generating a single omnidirectional panorama video stream. Although spatial and color calibration algorithms are embedded into the device to control the panorama generation, certain artifacts still remain.

The following sections review algorithms used to provide the general quality of the video, creating relations between the contribution of each single camera separately and the final panorama as a whole.

2.2.1. Detecting Blockiness

The methodology presented by Wang *et al.* in [21] has been used to detect blockiness in our test videos. The typical panorama video analyzed, generated by

stitching together the contribution of five different camera streams, is based on JPEG compression. JPEG is a block DCT-based lossy video coding technique. It is lossy because of the quantization operation applied to the DCT coefficients in each 8×8 coding block. Therefore, both blockiness and blurring artifacts may be created during the quantization process. Blocking effects occur due to the discontinuity at block boundaries, since JPEG quantization is block-based and the blocks are quantized independently. Blurring effects are mainly due to the loss of high frequency DCT coefficients, which smoothes the image signal within each block.

By transforming sampled images of the video into the frequency domain, we can effectively examine both blocking and blurring effects. As an example, let us denote the test image signal as $x(m, n)$ for $m \in [1, M]$ and $n \in [1, N]$, and let us calculate a difference signal along each horizontal line as

$$d_h(m, n) = x(m, n + 1) - x(m, n), \;\; n \in [1, N - 1]. \tag{2.1}$$

If we now let $f_m(n) = |d_h(m, n)|$ be a 1-D horizontal signal for a fixed value of m and we compute the power spectrum of $f_m(n)$ for $m = 1, 2, ..., M$, we can average them together to obtain a power spectrum estimation $P_h(l)$ like the one shown in Fig. 2.3. In this plot, the blocking effect can be easily identified by the peaks at the frequencies 1/8, 2/8, 3/8, and 4/8; the blurring effect is also characterized by the energy shifting from high frequency to low frequency bands.

A well-known disadvantage of the frequency domain method is the use of the Fast Fourier Transform (FFT), which has to be calculated many times for each image. The FFT also requires more storage space because it cannot be computed locally. The method proposed in [21] has been adopted because it can overcome this problem by providing a feature extraction procedure that is memory-efficient

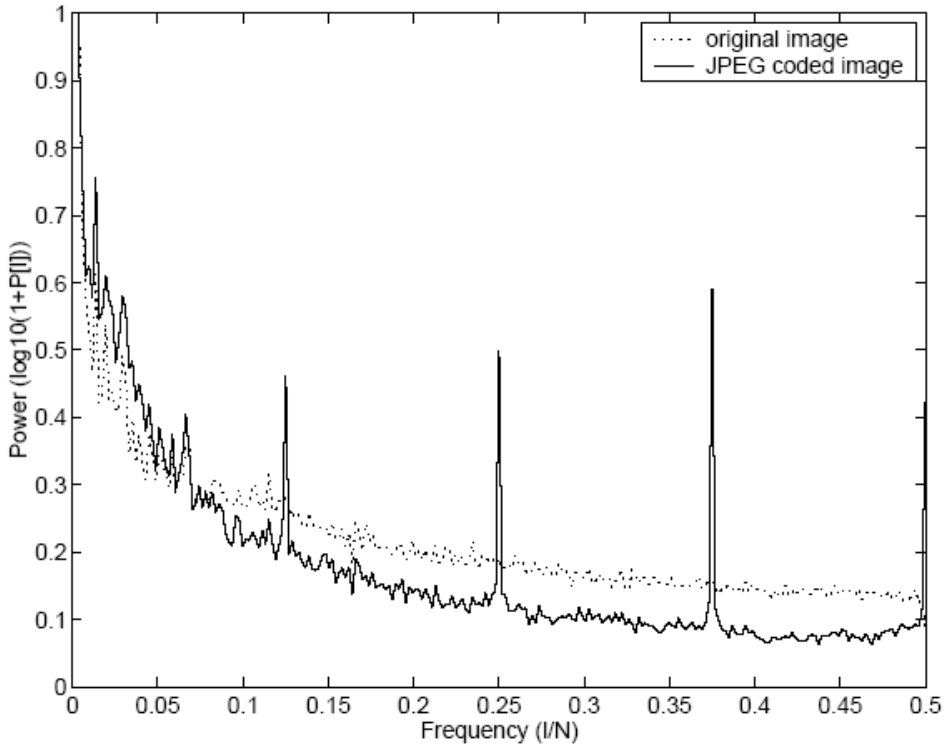

FIGURE 2.3. Power spectrum comparison between the original "Lena" image and its JPEG compressed version, [21].

and computationally inexpensive. A brief description of this method is given in the rest of this paragraph.

The features are calculated, using similar methods, first horizontally and then vertically. For convenience, only the horizontal feature extraction is presented. First, an estimation of the blockiness is obtained as the average difference across block boundaries:

$$B_h = \frac{1}{M(\lfloor N/8 \rfloor - 1)} \sum_{1=1}^{M} \sum_{j=1}^{\lfloor N/8 \rfloor - 1} |d_h(i, 8j)|, \tag{2.2}$$

where $[M, N]$ define the size of the image and $d_h(\cdot)$ is the difference signal defined in Eq. (2.1).

Second, the activity of the image signal is estimated. Blurring causes the reduction of signal activity, and combining the blockiness and activity measures gives more insight into the relative blur in the image. The activity is measured using two factors:

1. The first is the average absolute difference between in-block image samples:

$$A_h = \frac{1}{7}\left[\frac{8}{M(N-1)}\sum_{1=1}^{M}\sum_{j=1}^{N-1}|d_h(i,j)| - B_h\right] \tag{2.3}$$

2. The second activity measure is the zero-crossing (ZC) rate. For $n \in [i, N-2]$ we define

$$z_h(m,n) = \begin{cases} 1, \text{ horizontal } ZC \text{ at } d_h(m,n), \\ 0, \text{ otherwise.} \end{cases} \tag{2.4}$$

The horizontal ZC rate can then be estimated as

$$Z_h = \frac{1}{M(N-2)}\sum_{1=1}^{M}\sum_{j=1}^{N-2} z_h(m,n). \tag{2.5}$$

Using similar methods, we calculate the the vertical features of B_v, A_v, and Z_v. Finally, the overall features are given by

$$B = \frac{B_h + B_v}{2}, \;\; A = \frac{A_h + A_v}{2}, \;\; Z = \frac{Z_h + Z_v}{2}. \tag{2.6}$$

There are many different ways to combine the features to constitute a quality assessment model, but according to Wang *et al.* [1], one method with good prediction performance is the following

$$S = \alpha + \beta B^{\gamma_1} A^{\gamma_2} Z^{\gamma_3}\,, \tag{2.7}$$

where α, β, γ_1, γ_2, and γ_3 are model parameters that must be estimated with the subjective test data. The nonlinear regression routine *nlinfit.m* in the Matlab Statistics Toolbox is used to find the best parameters for Eq. (2.7). The model has performed well in all our tests, showing efficiency and robustness.

2.2.2. Detecting Blur

Beside the method presented in the previous paragraph, another NR perceptual metric has been taken into consideration to quantify blurring effects. Proposed by Marziliano *et al.* [22], the method is particularly interesting because it does not assume any knowledge of the original image and it does not make any assumption on the type of content or blurring process. Starting from the evidence that blur is perceptually noticeable along edges or textured areas, the technique is based on the smoothing effect of blur on edges and, consequently, attempts to measure the spread of the edges. In practice, the interesting result of the method is that measuring blur along vertical edges has been demonstrated to be sufficient. The algorithm can be outlined as follows:

1. An edge detector (e.g., a vertical Sobel filter) is applied to find the vertical edges in the image;

2. Each row of the image is scanned;

3. For pixels corresponding to the edge location, the start and end positions of the edge are defined as the local extrema locations closest to the edge;

4. The edge width is then calculated by subtracting the end position from the start position, and the local blur measure for this edge location is identified;

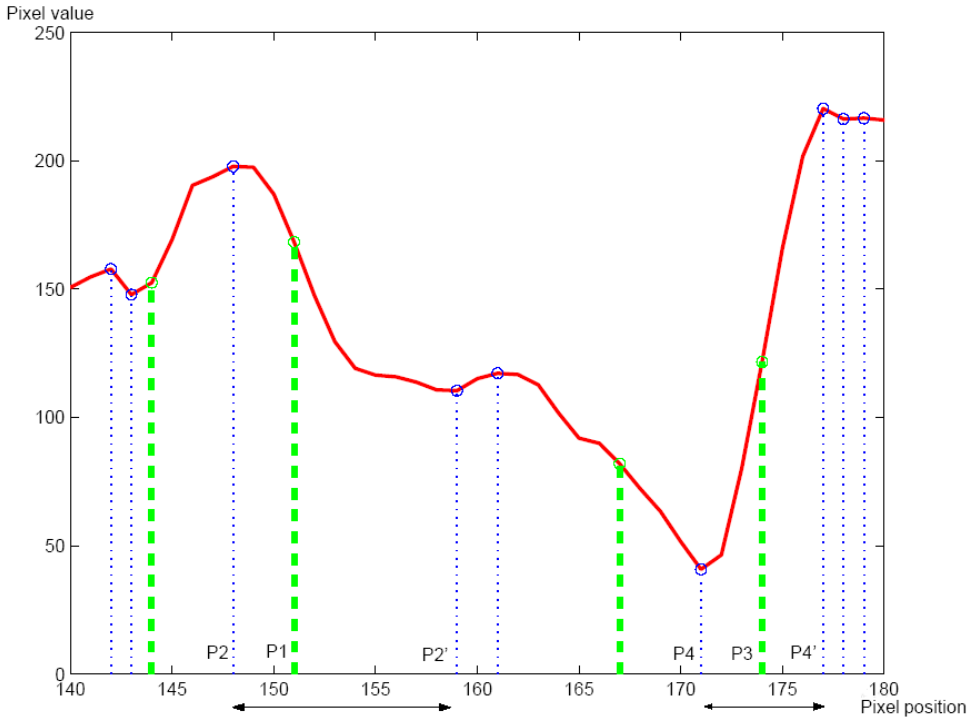

FIGURE 2.4. One row of the blurred image. The detected edges are indicated by the dashed lines, and local minima and maxima around the edge by dotted lines. The edge width at $P1$ is given by $P2' - P2$ [22].

5. Finally, the global blur measure for the whole image is obtained by averaging the local blur values over all edges locations.

Fig. 2.4 shows an example of the analysis of a row in a image (Y channel). For the edge location $P1$, the local maximum $P2$ defines the start position, while the local minimum $P2'$ corresponds to the end position. The edge width is $P2' - P2$. Similarly, for the edge $P3$, the local minimum $P4$ is the start position, the local maximum $P4'$ is the end position, and $P4' - P4$ is the edge width.

The method is particularly attractive because is near real-time, has shown low computational complexity, and its performance is independent of the image content.

2.2.3. Detecting Jerkiness and Noise

A modified structural distortion metric introduced by Wang *et al.* [1] is used on subsequent frames to identify regions that are supposed to maintain the same quality from a spatiotemporal point of view. The metric yields a mapping of the quality of these 'static' areas, identifying potential noise injection, jerkiness, and temporal fluctuations (flicker).

As mentioned in Section 2.1, many of the quality metrics currently available attempt to predict perceptual quality by quantifying perceptible errors. On the other hand, these techniques consider the "structural information" in an image as those characteristics that reflect the structure of the objects in the scene. This is independent of the average luminance and contrast of the image and leads to an image/video quality assessment approach that separates the measurement of luminance, contrast, and structural distortions. The luminance of the surface of an object being observed is the product of illumination and reflectance, but the structure of the objects in the scene is independent of illumination. Consequently, to explore the structural information in an image, the influence of the illumination is separated.

Suppose that $\mathbf{x}$ and $\mathbf{y}$ are two nonnegative image signals, which have been aligned with each other (e.g., spatial patches extracted from each image). If we consider one of the signals to have flawless quality (say $\mathbf{x}$), then the similarity measure can serve as a quantitative measurement of the quality of the second

signal $\mathbf{y}$. Let μ_x,μ_y,σ_x^2,σ_y^2 and σ_{xy} be the mean of x, the mean of y, the variance of x, the variance of y, and the covariance of x and y, respectively. Here the mean and the standard deviation (square root of the variance) of a signal are roughly considered as estimates of the luminance and contrast of the signal. The covariance (normalized by the variance) can be thought of as the degree of linear correlation between $\mathbf{x}$ and $\mathbf{y}$. Therefore, the metric formally separates the task of similarity measurement into three comparisons: **luminance** and **contrast** distortion, and loss of similarity in **structure** (correlation). Luminance, contrast, and structure comparison measures can be defined as follows:

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2}, \quad c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}, \quad s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy}}{\sigma_x\sigma_y}. \tag{2.8}$$

It is worth noting that these terms are conceptually independent in the sense that the first two terms only depend on the luminance and the contrast of the two images being compared, respectively, and purely changing the luminance or the contrast of either image will not affect the third term, the structure. Geometrically, $s(\mathbf{x},\mathbf{y})$ corresponds to the cosine of the angle between vectors $\mathbf{x}-\mu_x$ and $\mathbf{y}-\mu_y$. Although $s(\mathbf{x},\mathbf{y})$ does not use a direct descriptive representation of the image structures, it reflects the similarity between two image structures by stating that it equals one if and only if the structures of the two image signals being compared are exactly the same (recall that we consider structural information as those image attributes other than the luminance and contrast information).

When $(\mu_x^2+\mu_y^2)(\sigma_x^2+\sigma_x^2) \neq 0$, the similarity index measure between $\mathbf{x}$ and $\mathbf{y}$ given in [7] corresponds to

$$S(\mathbf{x},\mathbf{y}) = l(\mathbf{x},\mathbf{y}) \cdot c(\mathbf{x},\mathbf{y}) \cdot s(\mathbf{x},\mathbf{y}) = \frac{4\mu_x\mu_y\sigma_{xy}}{(\mu_x^2+\mu_y^2)(\sigma_x^2+\sigma_y^2)}. \tag{2.9}$$

If the two signals are represented discretely then the statistical features can be estimated as follows:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{2.10}$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2, \quad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2\,, \tag{2.11}$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y). \tag{2.12}$$

Notice that, especially over flat regions, an instability problem may arise if, in Eq. (2.9), $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is close to 0. In order to avoid this problem, a modification is necessary in the definitions of Eq. (2.8); in particular:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{2.13}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{2.14}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{2.15}$$

where $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, L is the dynamic range of the pixel values ($L = 255$ for 8-bit/pixel gray-scale images), and K_1 and K_2 are two constants whose values have to be small and such that C_1 or C_2 have effect only when $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is small. Indicative values for K_1 and K_2 are $K_1 = 0.01$ and $K_2 = 0.03$, respectively. For convenience, in this work we set $C_3 = C_2/2$. The resulting new measure is named the Structural SIMilarity (SSIM) index between signals $\mathbf{x}$ and $\mathbf{y}$:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{2.16}$$

A more generalized version of the resulting similarity measure takes into account also potential different contributions of the three terms of the comparison (luminance, contrast, structure).

$$S(\mathbf{x}, \mathbf{y}) = \alpha[l(\mathbf{x}, \mathbf{y})] \cdot \beta[c(\mathbf{x}, \mathbf{y})] \cdot \gamma[s(\mathbf{x}, \mathbf{y})] \tag{2.17}$$

where α, β, and γ are parameters used to adjust the relative importance of the three components. Setting $\alpha = \beta = \gamma = 1$, we obtained the simplified version of Eq. (2.16).

3. CHAPTER THREE

3.1. Quality Assessment of Panorama Video

Analyzing the results obtained applying standard techniques (such the ones presented in the previous chapter) to assess the quality of panorama video streams, it is immediately evident that the approach is not satisfactory. Evidently, current quality assessment techniques are not adequate and need to address specific problems. The overall quality of a panorama video stream as assessed with standard methods presents strong deviations from the results of subjective quality tests.

In particular, the following limitations have been encountered working with standard quality assessment methods:

1. Blockiness or noise confined to limited regions or single cameras; standard methods tend to average out the error and the overall quality prediction is excessively optimistic.

2. The differences in each camera calibration are not appropriately handled because not detectable by a general method, even though it is easily noticed by the human eye.

3. Camera vignetting and blooming may systematically occur in the upper or lower part of the panorama, affecting the overall quality of the video.

4. There exist no test vectors or reference panoramic video databases.

5. There exists no specific semantic level that addresses the evaluation of a video characterized by a large aspect ratio (i.e., the height much smaller than the width). The user focus of attention can have sweeping ranges,

be attracted by noticeable seam steps, and be mainly driven by motion tracking.

In other words, in addition to common artifacts present in digital video streams, new impairments introduced by the specific technology employed have to be detected. If the former are primarily due to compression and network conditions as we have seen in the previous chapter, the latter are mainly the result of combining the video signals of different sources into one single multimedia stream. The results have proved that this process seriously affect the MOS prediction usually obtained with standard objective metrics.

The idea is to first use and customize existing techniques to frame the general quality of the video (creating relations between the contribution of each camera separately and the final panorama as a whole) and then add several weighting stages to the metric in order to tackle the specific measurement deficiencies. A semantic level is appended to the quality measure to handle those cases in which cognitive behavior plays a determinant role. The pooling process is then driven by the output by assigning higher weight to the regions with numerically higher deviations from subjective scores and semantically higher significance. In this way, excessively optimistic results are bounded and, with the adjustment introduced, we produce a stronger prediction performance.

This Chapter is divided in four Sections. Section 3.2 discusses the adjusted quality metric for panorama video streams. Section 3.3 reports on some experimental results we have conducted to test our quality assessment system. Section 3.4 finally comments on the performance evaluation of the proposed metric.

3.2. Omnidirectional Video Quality Metrics

Low-level aspects of vision such as the contrast sensitivity (enhanced also by potentially strong steps between adjacent cameras), color perception (smearing, vignetting), masking, and general impairments are very important in video/image quality assessment. But the cognitive behavior of people when watching video sequences has to be taken in consideration as well, as argued by Cavallaro *et al.* [23] and Bajcsy [24]. A generalization is almost impossible given the large variety of behaviors generated in individuals by similar situations; however, some important characteristics can be extracted. Several researchers have demonstrated that humans are generally interested in what they look at [25, 26] and that the focus of the viewers strongly depends on the scene [27]. A spatiotemporal 'importance map' was proposed in [28] to determine a prediction of the focus of attention.

Tests have demonstrated that, even without considering more sophisticated approaches that take into account changes in the human visual system sensitivity (e.g., the loss of spatial acuity for the background when following a movement [5]) the existing correlation between motion localization/quantification and perceived quality is able, along with the seam analysis, to produce key weighting adjustments to the numerical analysis. As shown in Fig. 3.1, motion is quantified in each camera to detect static regions for noise analysis and, at the same time, create a first ranking of camera contribution to the semantic level. The range of the motion is then calculated in order to have a more precise correlation with the potential range of the focus of attention (see Fig. 3.2).

In other words, the motion information is used as a semantic to segment moving objects and create a spatio-temporal map of potentially impaired regions that will define the range of the focus of attention and will constitute, along

with impaired seam areas, the likely target areas of judgment. A high-level segmentation is thus added in our system to check where the motion occurs and to appropriately identify regions that can potentially attract the focus of attention.

In the following sections, we present and discuss each component that is used to build our global quality metric.

3.2.1. High-Level Vision Factors in Quality Assessment

3.2.1.1. Motion tracking and quantification

A modified structural distortion metric introduced by Wang *et al.* [1] is used on subsequent frames to identify regions of the panorama where the motion is concentrated. At the same time, this metric yields a mapping of the quality of the areas that are supposed to maintain the same quality from a temporal point of view. This method is also useful for automatically detecting temporal fluctuations (flicker) problems. The regions where the motion occurs are marked as potentially significant because it is where blockiness and blurriness problems may arise and where the observer's focus of attention is more likely concentrate. Since in the case under investigation the scenarios are in most cases limited to conference rooms and offices, a strong correlation exists between the person talking and the focus of attention of the consumer.

Therefore, the observer, as expected, tends to follow the human action. From a perceptual point of view, this suggests the creation of a ranking of each camera contribution based on the presence of animated subjects and, at the same time, the determination of the areas where this is more likely to occur in each camera contribution, tracking the changes both spatially and temporally.

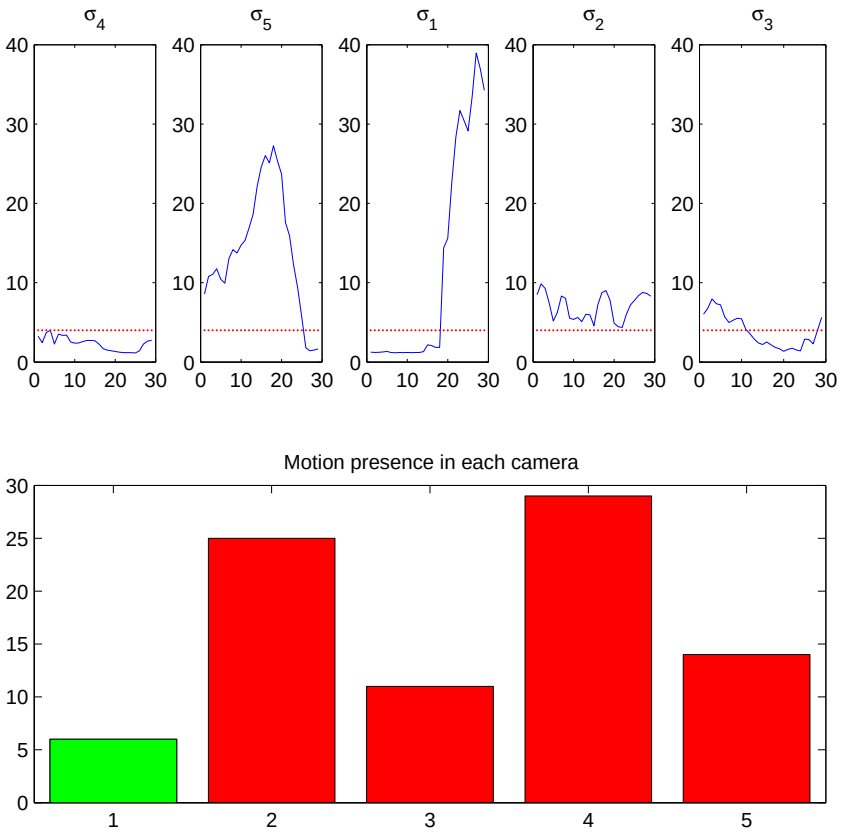

FIGURE 3.1. Motion quantification analysis. The bars indicate the amount of motion occurring in each camera. The magnitude of each bar is related to the behavior of the correspondent standard deviation σ_n calculated over subsequent frames.

3.2.1.2. Seams regions

Regions at the seam between adjacent cameras are analyzed mainly to identify calibration and vignetting impairments, introduced by stitching images from different cameras with similar, but not equivalent, calibrations and characterizations. To achieve a seamless panorama, a general color calibration is usually implemented. Absolute color calibration is generally avoided since it is difficult to provide a reference source spanning all lighting conditions. Relative

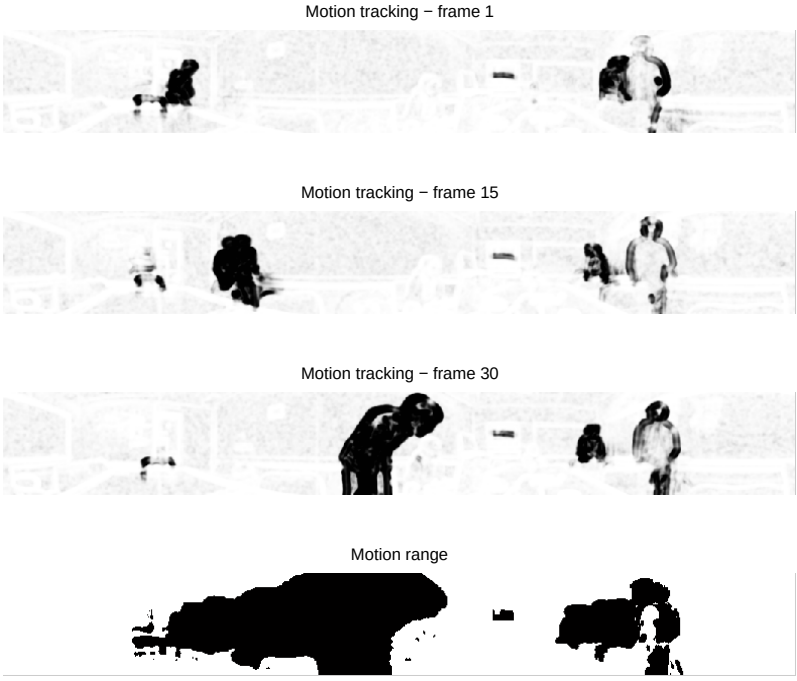

FIGURE 3.2. Example of spatio-temporal map of potentially impaired regions and focus of attention concentration. The first, second, and third image represent the motion detected at three different times, while the fourth image represents the entire motion range detected.

calibration is usually adopted instead. Between one and four columns in the region of two camera edges (seam region) are sampled and a transfer function is generated to compare the mean and standard deviation of each adjacent camera image to the mean and the standard deviation of this region. Even not assuming an affine/linear correction, as in the devices used for our tests, the method has shown consistent results.

Furthermore, as shown in Fig. 3.3, possible steps in the luminance channel of adjacent cameras images are estimated using sampling lines over multiple

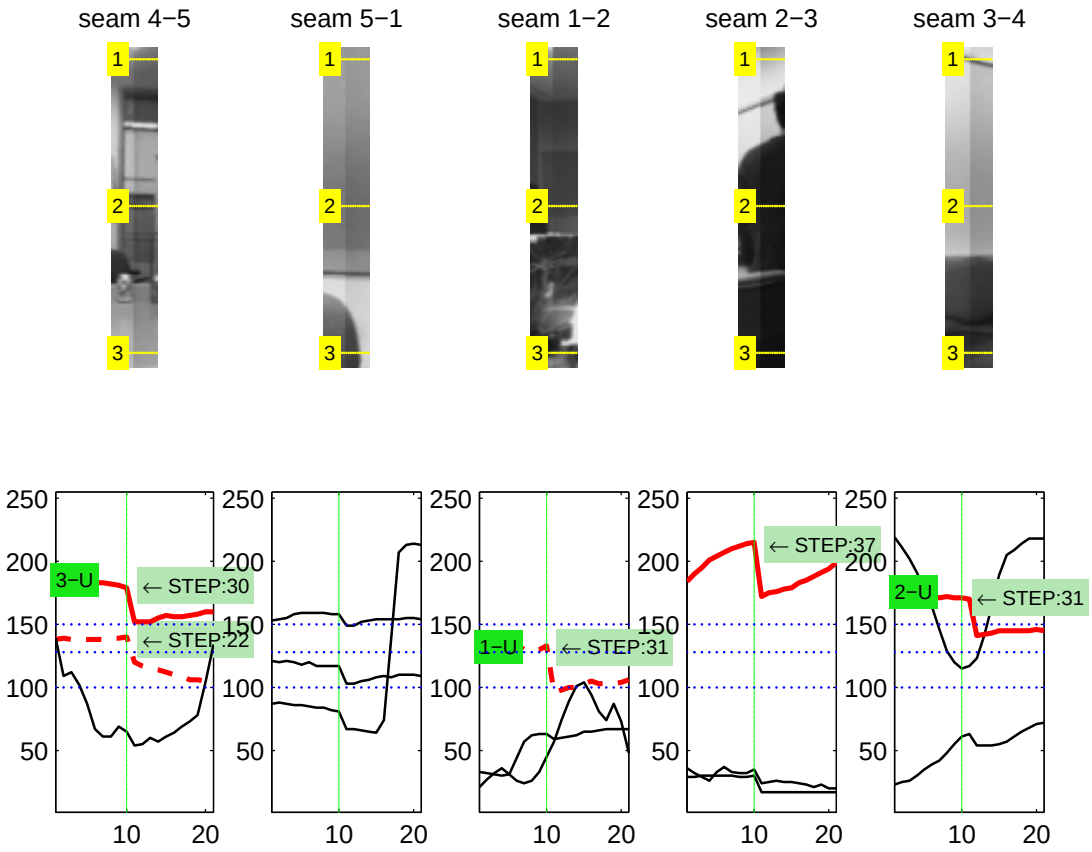

FIGURE 3.3. Seam region analysis. Luminance level, magnitude of luminance steps, and uniform regions are analyzed.

frames. The contribution of the steps is associated with camera pairs and ranked according to their magnitude and the luminance level where they occur (bold red lines and dashed lines denote larger steps and occurring at higher luminance levels, while black lines denote less visible regions or lower steps).

It is in fact well-known that the human eye is very sensitive to overall intensity (luminance) changes and subjective test have demonstrated that the magnitude and the level at which the step occurs play an important role in attracting the focus of attention and, consequently, in weighting appropriately the

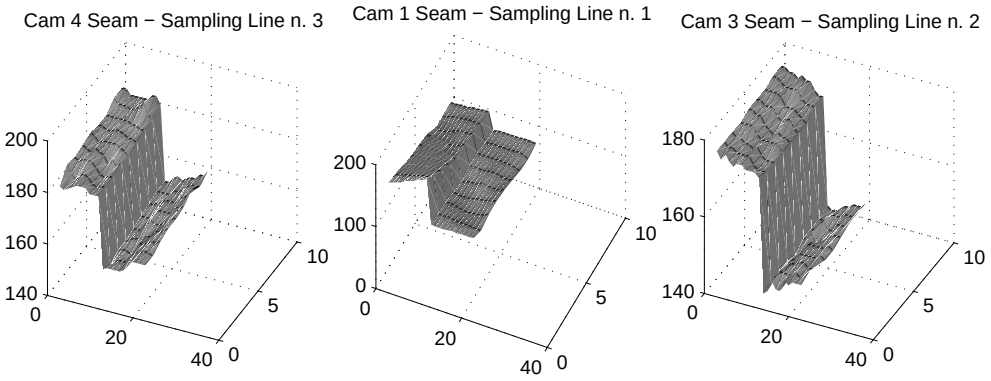

FIGURE 3.4. Uniform areas are automatically extracted in proximity of a calibration impaired seam (large step) between two cameras.

user's perceived quality. Moreover, by analyzing the slope of the luminance in multiple sampling lines close to the seam region, it is possible to identify uniform or quasi-uniform regions where one can make a consistent comparison between samples of the same surface lying in the field of view of two adjacent cameras.

We adopt a tolerance factor that takes into account the variation of the illumination that may occur in the sampled surface. By using an algorithm that compares the entire region of the seam, we can easily identify possible objects lying exactly at the boundary between two cameras. Two patches from correspondent 'static' regions can thus be extracted and analyzed in order to obtain a relative measure of the noise occurring temporally in that region and an estimate of the luminance step. The top and bottom regions of the panorama near the seam are also analyzed to detect potential vignetting problems (anomalous radial drops of the intensity profile). Fig. 3.4 shows uniform areas around the seam step extracted in proximity of a seam between two cameras.

3.2.2. Low-Level Vision Factors in Quality Assessment

From a purely mathematical point of view, the no-reference quality metric estimates visual quality based on the analysis of three main video artifacts: blockiness [29], blur [22], and jerkiness. The perceptual metrics presented in the previous chapter are applied locally to each camera contribution over a predefined number of frames in each video stream. The overall quality rating is driven by the result of the semantic segmentation based on the seam uniformity and motion quantification/tracking discussed above.

3.3. Experiments and results

Since a standard database of test panoramic videos is not available yet, we have generated a panoramic video database which covers a variety of scenarios with different lighting conditions, people of different gender and ethnicity, different motion patterns. Our proposed perceptual metric has been evaluated by performing sessions of subjective testing, conveniently divided in three phases: instruction/training, main test, discussion. Some details on the subjective test procedure is given in the following paragraphs.

3.3.1. Subjective Evaluation of Video Quality

Subjective evaluation experiments are complicated in many respects. Viewing conditions and human psychology, for instance, are two key factors that can heavily affect a subjective evaluation. Other factors include observer vision ability, translation of quality perception into ranking score, preference for content, adaptation, display devices, ambient light levels, just to name a few.

Both MPEG [35, 36] and ITU [30–33] have made recommendations on how to perform subjective tests for evaluating the quality of digital video images coded at low and medium bit rates. These tests methods are based on traditional subjective test methods, although adapted to address specific characteristics of the MPEG-4 encoding scheme (low target bit rates, channel error conditions, etc.). The two methods that we will briefly present are the Single Stimulus Continuous Quality Evaluation (SSCQE) and the Double Stimulus Continuous Quality Scale (DSCQS). These methods have demonstrated to have repeatable and stable results and have consequently been adopted as a part of the international standard by the ITU. In our case the SSCQE and DSCQS tests (simulated and for evaluators training purposes) have been conducted on multiple subjects and the scores have been averaged to yield the MOS. The standard deviation between the scores may also be useful to measure the consistency across different subjects.

3.3.1.1. Single Stimulus Continuous Quality Evaluation

In the SSCQE method, subjects indicate their impression of the video quality on a linear scale that is divided into five segments. The five intervals are marked with adjectives which serve as guides. In our tests, six additional descriptive words are made available to the evaluator to help the evaluation process: overall video quality, visible artifacts, strong impairments, noise, color quality, panorama uniformity over 360 degrees. The subjects are instructed to mark any point on the scale that best reflects their impression of quality at that time instant, and to track the changes in the quality of the video using a slider. This method is appropriate when references are not available as in our case.

3.3.1.2. Double Stimulus Continuous Quality Scale

The DSCQS method is a form of discrimination-based method and offers the extra advantage that the subjective scores are less affected by adaptation and contextual effects. In the DSCQS method, the reference and the distorted videos are presented one after the other in the same session, in small segments of a few seconds each, and subjects evaluate both sequences using sliders similar to those used for the SSCQE method. The difference between the scores of the reference and the distorted sequences gives the subjective impairment judgement. In our NR scenario, this test has been conducted with specific synthetically impaired video stream to train the observers and have consistency with single stimulus continuous quality evaluation. A network emulator, for instance, has been used to recreate specific network impairment in the video and analyze its impact on the quality evaluation of the observers.

3.3.2. Quality Assessment and Prediction Performance

The quality assessment with the new method can be summarized as follow: First, blockiness and blur for each camera stream are estimated by averaging the results obtained with the two methods presented in Chapter 2. We can define the blur/blockiness quality coefficient as follows

$$q_i(S_i, b_i) = \frac{1}{2}(S_i + b_i) = \frac{1}{2}(\alpha + \beta B_i^{\gamma_1} A_i^{\gamma_2} Z_i^{\gamma_3} + b_i), \tag{3.1}$$

where $i = 1, ..., 5$ is the number of cameras, S_i is the output of the blur/blockiness metric defined in Eq. (2.7), and b_i is the output of the blur metric presented in Section 2.2.2.

Second, the motion quantification weight $W_i^{M_q}$ and the motion tracking weight $W_i^{M_T}$ are calculated. Motion quantification is obtained by applying the structural distortion method, as defined in Eq. (2.16), to subsequent frames. Each camera contribution is ranked according to the variation over time of the standard deviation against a given threshold. Cameras with a larger motion presence are assigned a larger weight. Using the same approach, motion tracking is calculated to define a hypothetical range for the focus of attention. Cameras with regions that fall into the motion range are weighted more.

Third, seam analysis is performed on the stitching regions of the panorama. A fixed number of rows in these regions is scanned. The perceptible steps are identified by evaluating their magnitude and the luminance level where they occur. By combining the contributions of these two factors, we can define the seam quality coefficient as follows

$$W_i^S = \frac{1}{N} \sum_{n=1}^{N} m_n \cdot l_n \, , \tag{3.2}$$

where $i = 1, ..., 5$ is the number of the camera whose rightmost seam is under investigation, N is the number of perceptible steps, and m_n and l_n are the magnitude and the luminance level coefficient of the *n-th* step, respectively. The video streams from cameras with larger steps occurring at higher luminance levels (i.e., more visible) are weighted more.

A Video Quality score for the panorama is then generated by combining the weighted contributions of each camera as

$$VQ = \sum_i W_i^{M_q} W_i^{M_t} W_i^S q_i(S_i, b_i). \tag{3.3}$$

Finally, a predicted MOS associated with the VQ score is generated (using non-linear regression as suggested in [37]) and correlated with the MOS score given to the same video through subjective tests.

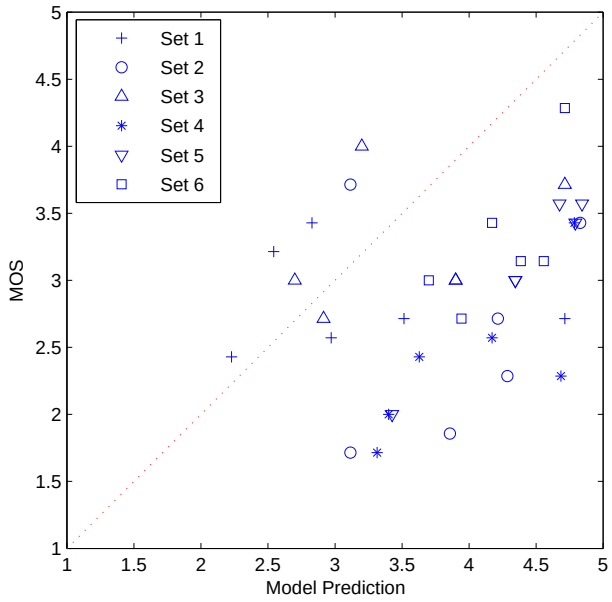

FIGURE 3.5. Scatter plots of MOS versus not adjusted model prediction for the six video datasets tested.

To assess the enhancement of the prediction performance attributable to seam and motion segmentation, we make a comparison between the results obtained with standard perceptual metric analysis (see Fig. 3.5) and those obtained with the adjusted metric proposed (see Fig. 3.6). As can be seen in these figures, the modified metric in general gives better results, achieving a stronger correlation between subjective tests and metric's predictions. As expected, the improvement is particularly significant when dealing with video impaired by evident calibration problems and by localized artifacts in limited regions of the panorama. In these cases, standard methods usually give too optimistic predictions.

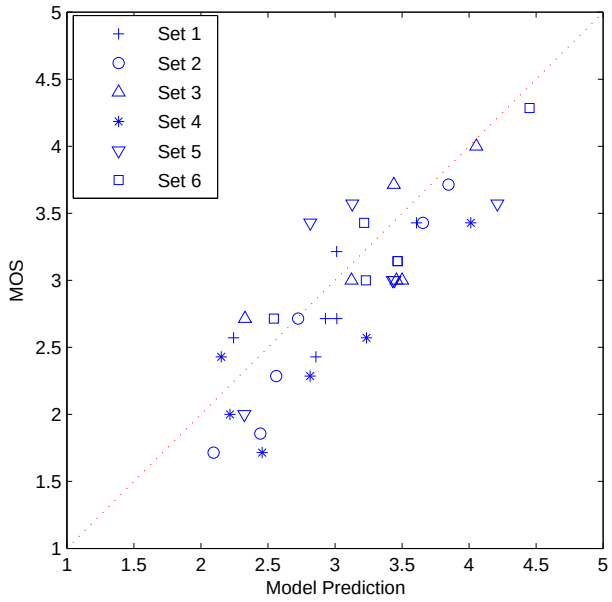

FIGURE 3.6. Scatter plots of MOS versus adjusted model prediction for the six video datasets tested.

3.4. Performance Evaluation

The new objective quality metric has to prove to be a reliable, repeatable, and cost-effective measure for video quality assessment applications. The goal of this objective model is to predict perceived video quality and to find a consistent correlation with the subjective evaluation. Therefore, it is mandatory to build a video database covering numerous scenarios and to have a subjective evaluation score associated with each entry. Such a database can be used to assess the prediction performance to validate the objective quality measurement algorithms.

A systematic way to evaluate the prediction performance of the objective model is to measure three specific attributes [31]:

- *Prediction Accuracy.* Prediction accuracy of an objective model is the ability of the model to predict subjective quality ratings with minimum error.

- *Prediction Monotonicity.* MOS predicted values of the objective model should be monotonic over the range of the corresponding subjective MOS values. The objective model was evaluated for its prediction ability by analyzing how the predicted MOS values behaved with the varying subjective results.

- *Prediction Consistency.* Prediction consistency is the ability of the model to provide consistent and accurate predictions for the data from the subjective tests on a large range of video sequences.

The metrics used to measure these prediction attributes are presented in the following paragraphs.

3.4.1. Pearson Correlation Coefficient

The Pearson correlation coefficient was used to model the prediction accuracy of the proposed video quality metric. The Pearson correlation coefficient is a statistical measure of the correlation between two sets of data (i.e., the objective data set and the subjective data set). The range of the correlation is between -1 and 1. A correlation figure of 1 represents the largest similarity between two variables, 0 means that there is no correlation, and a correlation value of -1 indicates a perfect negative correlation. The prediction accuracy was computed by correlating the predicted MOS of the objective model with the MOS of the subjective tests [32, 33]. The formula for calculating the Pearson correlation coefficient is given by

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}\,, \tag{3.4}$$

where x_i and y_i, $i = 1, ..., N$, are the sets of data from the subjective and the objective models which are being compared for similarity, and N is the size of the data sets.

Table 3.1 shows the values of the Pearson correlation for the data of the new method data and a standard perceptual metric. While the standard method reveals negative or low values, those for the new metric are close to 1. The proposed metric therefore achieves higher prediction accuracy than a standard perceptual metric.

TABLE 3.1. Pearson Correlation Coefficients.

Video Sequence	**Standard perceptual metric**	**Adjusted metric**
Video 1	-0.5873	0.8732
Video 2	-0.2883	0.8114
Video 3	0.4315	0.9326
Video 4	-0.4221	0.8553
Video 5	-0.7234	0.8605
Video 6	0.5234	0.9705

3.4.2. Spearman Rank Correlation

Spearman's correlation was used as a measure of prediction monotonicity between the subjective MOS and the predicted MOS. This method is a quantitative measure of the strength of a relationship between two sets of data [31]. The range of the Spearman rank correlation is between -1 and 1, where a correlation

of 1 indicates the most monotonic relationship. Spearman rank correlation was calculated using the following equation:

$$\rho_{XY} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (X_i - Y_i)^2 \,, \tag{3.5}$$

where X_i is the rank of x_i and Y_i is the rank of y_i in the ordered data series. As illustrated in Table 3.2, the values of the Spearman correlation for the proposed metric shows a strong monotonic relationship with the subjective results as compared to the same relationship of a standard perceptual metric. The values prove that, unlike the standard metric, the proposed metric is monotonic in the range of the subjective data for all video sequences considered.

TABLE 3.2. Spearman Rank Correlation Coefficients.

Video Sequence	**Standard perceptual metric**	**Adjusted metric**
Video 1	-0.3563	0.9231
Video 2	-0.1554	0.9321
Video 3	0.5215	0.9743
Video 4	-0.3113	0.8921
Video 5	-0.6562	0.9283
Video 6	0.3981	0.9678

3.4.3. Outlier Ratio

The number of outliers with respect to the total number of datapoints was used as a measure of prediction consistency between the subjective MOS and the predicted MOS. The threshold for defining an outlier point was set at twice the MOS Standard error.

4. CHAPTER FOUR

4.1. Conclusions

The very limited literature currently available on quality assessment of panorama video has motivated us to work on a new metric to address specific impairments of this type of video streams. Several standard quality assessment methods were introduced and used to compute the general quality of each camera video separately. Relations have then been established between the single video contributions and the panorama video, obtained by stitching independent video streams together.

The topic of panorama video quality assessment was explored starting from specific limitations encountered using standard quality assessment approaches and deriving several weighting stages in order to tackle specific measurement deficiencies. The weighting functions added to the standard approaches led to a high correlation with the subjective ratings and effectively captured those impairments that were affecting the subjective/objective correlation the most.

Important factors for a reliable panorama video quality assessment were introduced, such as the seam analysis and the motion quantification/tracking segmentation as parts of a semantic level. The semantic level was appended to the quality measure to efficiently handle those cases in which cognitive behavior plays a determinant role.

Performance comparisons between the proposed method and standard approaches were presented.

A Matlab Video Quality Tool for objective panorama video quality assessment presented in this thesis was developed.

4.2. Future Work

With the possibility to have source and destination video recorded, future research will include a FR panorama quality metric study and implementation.

An integration of a simplified version of the strategies presented is under investigation to monitor and optimize the quality of the panorama at the generation stage.

How to quantify and incorporate color distortions also needs more research efforts.

BIBLIOGRAPHY

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image Quality Assessment: From Error Measurementes to Structural Similarity," *IEEE Transactions On Image Processing*, Vol. 13, No. X, Jan. 2004, pp. 325-376.

[2] D.A. Silverstein, J.E. Farrell, "The relationship between image fidelity and image quality," in *Proc. of the International Conference on Image Processing (ICIP)*, Lausanne, Switzerland, 1996.

[3] Z. Wang, L. Lu, A. C. Bovik, "Video Quality assessment Based on Structural Distortion Measurement," *Signal Processing: Image Communication*, Vol. 19, No. 2, Feb. 2004, pp. 121-132.

[4] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Processing*, pp. 982-986, 1994.

[5] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, pp. 177-200, Nov. 1998.

[6] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Communications*, vol. 43, pp. 2959-2965, Dec. 1995.

[7] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81-84, March 2002.

[8] Z. Wang, A. C. Bovik and L. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 4, pp. 3313-3316, May 2002.

[9] J.-B. Martens and L. Meesters, "Image dissimilarity," *Signal Processing*, vol. 70, pp. 1164-1175, Aug. 1997.

[10] ANSI T1.801.01-1996, "Digital Transport of Video Teleconferencing / Video Telephony Signals - Video Test Scenes for Subjective and Objective Performance Assessment," American National Standards Institute, 1996.

[11] ANSI T1.801.02-1996, "Digital Transport of Video Teleconferencing / Video Telephony Signals - Performance Terms, Definitions, and Examples," American National Standards Institute, 1996.

[12] ANSI T1.801.03-1996 "Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment," American National Standards Institute, 1996.

[13] A. C. Bovik, S. Liu, "DCT-domain blind measurement of blocking artifacts in DCT-coded images" in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 3, pp. 1725-1728, May 2001.

[14] P. Gastaldo, S. Rovetta and R. Zunino, "Objective assessment of MPEG-video quality: a neural-network approach," in *Proc. IJCNN*, vol. 2, pp. 1432-1437, 2001.

[15] M. Knee, "A robust, efficient and accurate single-ended picture quality measure for MPEG-2," available at http://www-ext.crc.ca/vqeg/frames.html, 2001.

[16] H. R. Wu, M. Yuen, "A generalized block-edge impairment metric for video coding," in *IEEE Signal Processing Letters*, vol. 4, pp. 317-320, Nov. 1997.

[17] J. E. Caviedes, A. Drouot, A. Gesnot, and L. Rouvellou, "Impairment metrics for digital video and their role in objective quality assessment," in *Proc. SPIE*, vol. 4067, pp. 791-800, 2000.

[18] C. J.van den Branden Lambrecht, M. Kunt, "Characterization of human visual sensitivity for video imaging applications," in *ISignal Processing*, Vol. 67, pp. 255-69, 1998.

[19] D. S. Taubman, M. W. Marcellin, "JPEG2000: Image compression Fundamentals, Standards and Practice," Kluver Acadamic Publisherr, 2002.

[20] M. Yuen, H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions." in *Signal Processing*, Vol. 70, No. 3, pp 247 278, 1998.

[21] Z. Wang, H. R. Sheikh and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proceedings of IEEE 2002 International Conferencing on Image Processing*, Rochester, NY, September 22-25, 2002.

[22] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A No-Reference Perceptual Blur Metric," in *Proc. International Conference on Image Processing*, Rochester, NY, Sept. 2002.

[23] A. Cavallaro, S. Winkler, "Segmentation-driven perceptual quality metric," in *Proc. of IEEE Int. Conference on Image Processing (ICIP)*, Singapore, Oct. 2004.

[24] R. Bajcsy, "Active perception," in *Proc. of IEEE*, Vol. 76, n. 8, pp. 996-1005, 1988.

[25] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye Movements and Vision*, L. Riggs Editor, , pages 171-196, Plenum Press, New York, 1967.

[26] P. Barber and D. Legge, "Perception and Information," in *Information Acquisition*, Chapter 4, Methuen, London, 1976.

[27] L. B. Stelmach and W. J. Tam, "Processing image sequeences based on eye movements," in *Proc. SPIE*, Vol. 2179, pp. 90-98, San Jose, CA, 1994.

[28] A. Maeder, J. Diederich, E. Niebur, "Limiting human perception for image sequences," in *Proc. SPIE*, Vol. 2657, pp. 330-337, San Jose, CA, 1996.

[29] Z. Wang, A. C. Bovik, B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. ICIP*, Vol. 3, pp. 981-984, Vancouver, Canada, 2000.

[30] ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," in *International Telecommunication Union*, Geneva, Switzerland, 2002.

[31] ITU-R Document 6Q/14, "Final Report from the Video Quality Expert Group on the Validation of Objective Models of Video Quality Assessment," Phase II (FR-TV2), Sept. 2003.

[32] ITU-T Recommendation P.910, Subjective video quality assessment methods for multimedia application, International Telecommunication Union, 1996.

[33] ITU-T Recommendation P.920, Interactive test methods for audiovisual communications, International Telecommunication Union, 1996.

[34] ITU-T Recommendation P.930, Principles of a reference impairment system for video, International Telecommunication Union, 1996.

[35] T. Alpert, "Subjective Evaluation of MPEG-4 Video Codec Proposal: Methodological Approach and Test Procedures," 1995.

[36] ISO-IEC N1442, "Evaluation methods and procedures for July 97 MPEG-4 tests," International Organization for Standardization ISO-IEC/JTC1/SC29, WG11, 1996.

[37] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," http://www.vqeg.org/, Mar. 2000.

[38] C. J. van den Branden Lambrecht, D. M. Costantini, G. L. Sicuranza, and M. Kunt, "Quality assessment of motion rendition in videocoding," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 9, pp. 766-782. Aug. 1999.

[39] D. J. Heeger and T. C. Teo, "A model of perceptual image fidelity," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 343-345, 1995.

[40] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, A. B. Watson, ed., pp. 207-220, MIT Press, 1993.

[41] S. Winkler, "A perceputal distortion metric for digital color video," *Proc. SPIE*, vol. 3644, pp. 175-184, 1999.

[42] Y. K. Lai and C.-C. J. Kuo, "A Haar wavelet approach to compressed image quality measurement," *Journal of Visual Communication and Image Understanding*, vol. 11, pp. 17-40, Mar. 2000.

[43] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Information Theory*, vol. 4, pp. 525-536, 1974.

[44] A. B. Watson, "The cortex transform: rapid computation of simulated neural images," *Computer Vision, Graphics, and Image Processing*, vol. 39, pp. 311-327, 1987.

[45] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *Journal of Optical Society of America*, vol. 14, no. 9, pp. 2379-2391, 1997.

[46] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision* (A. B. Watson, ed.), pp. 163-178, Cambridge, Massachusetts: The MIT Press, 1993.

[47] J. Lubin, "A visual discrimination model for image system design and evaluation," in *Visual Models for Target Detection and Recognition*, E. Peli, ed., pp. 207-220, Singapore: World Scientific Publisher, 1995.

[48] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision* (A. B. Watson, ed.) pp. 179-206, Cambridge, Massachusetts: The MIT Press, 1993.

[49] J. Malo, R. Navarro, I. Epifanio, F. Ferri, and J. M. Artifas, "Non-linear invertible representation for joint statistical and perceptual feature decorrelation," in *Lecture Notes on Computer Science*, vol. 1876, pp. 658667, 2000.

[50] I. Epifanio, J. Gutirrez, and J. Malo, "Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding,' in *Pattern Recognition*, vol. 36, pp. 1679 1923, Aug. 2003.

APPENDICES

APPENDIX A. ANSI Scalar Quality Parameters

As part of the quality features and parameters suggested by ANSI, the Scalar Quality Parameters are based on Spatial Information SI and Temporal Information TI (see Chapter 1 for the definitions) extracted from the source and processed video.

In this Appendix, some widely used quality parameters based on TI and SI will be presented.

A.1. Scalar quality parameter based on temporal information

In this section some quality parameters based on the $TI(n)$ feature are introduced. According to the ANSI paper [8], the parameters $TI_S(n)$ and $TI_D(n)$ are preferably selected as $TI_S(n) = TI_{S,rms}(n)$ and $TI_D(n) = TI_{D,rms}(n)$, where S and D stand for Source and Destination, respectively. We can also use $TI_S(n) = TI_{S,mean}(n)$ and $TI_D(n) = TI_{D,mean}(n)$. This reduces the computational complexity but might also cause some degradation in accuracy of the computed parameters.

Parameter P1, Max of TI Ratio

$$TI_{ratio}(n) = \log 10\Big(\frac{TI_D(n)}{TI_S(n)}\Big) \tag{A.1}$$

$$P1 = \max[\max_{time} TI_{ratio}(n), 0] \tag{A.2}$$

If $TI_{ratio}(n)$ is positive, motion energy has been added to the destination frame compared to the source frame. As discussed above, this can occur as a result of added noise, jerkiness or error blocks. Negative values can occur as results

of frame repetition (i,e., lost motion).

Parameter P2, RMS of TI Ratio

$$P2 = RMS_{time}(TI_{ratio}(n)) = \frac{1}{N}\sum_{n} TI^2_{ratio}(n), \tag{A.3}$$

where N is the number of frames.

An average measure of how different the motion in the source sequence is from the destination sequence can be obtained by computing the $P2$ parameter.

Parameter P3, Max-Min of TI Ratio

$$P3 = \max[\max_{time}(TI_{ratio}(n)), 0] - \min[\min_{time}(TI_{ratio}(n), 0)] \tag{A.4}$$

P3 is similar to P1 but also includes lost motion energy (e.g., frame repetition).

Parameter P4, Positive Mean - Negative Mean of TI Ratio

$$P4 = \mu^{+}_{time}[TI_{ratio}(n)] - \mu^{-}_{time}[TI_{ratio}(n)] \tag{A.5}$$

where μ_{+} and μ_{-} are the mean of the positive and negative values respectively. The P4 parameter resembles the P3 paramenter but the former is less sensitive to peak values.

Parameter P5, RMS of TI Error Ratio(n)

$$TI_{errorratio}(n) = \frac{TI_S(n) - TI_D(n)}{TI_S(n)} \tag{A.6}$$

$$P5 = RMS_{time}(TI_{errorratio})(n) = \frac{1}{N}\sum_{n} TI^2_{errorratio}(n) \tag{A.7}$$

where N is the number of frames. The $TI_{errorratio}(n)$ is the difference in temporal information between the source and destination scene, normalized by the temporal information of the source.

Parameter P6, RMS of positive part of TI Error Ratio (lost motion energy)

$$P6 = RMS_{time}[\max(TI_{errorratio})(n), 0] \tag{A.8}$$

P6 provides an average measure of the lost motion energy.

A.2. Scalar quality parameter based on spatial information

In this section some quality parameters based on spatial information are introduced. According to the ANSI paper [8], the parameters $SI_S(n)$ and $SI_D(n)$ are preferably selected as $SI_S(n) = SI_S stdv(n)$ and $SI_D(n) = SI_D stdv(n)$. S and D stand for source and destination, respectively.

Parameter P7, Max Absolute Value of SI Error Ratio

$$P7 = RMS_{time}[\max(TI_{errorratio})(n), 0] \tag{A.9}$$

The $SI_{errorratio}(n)$ measures lost spatial information in the destination sequence compared to the source sequence. A positive value (lost SI) can occur as result of impairments like blurring, and a negative value as result of added noise, edges or error blocks. The parameter P7 is simply the maximum absolute value of $SI_{errorratio}(n)$.

Parameter P8, RMS of SI Error Ratio(n)

$$P8 = RMS_{time}(SI_{errorratio})(n) = \frac{1}{N}\sum_{n} SI^2_{errorratio}(n) \tag{A.10}$$

where N is the number of frames. P8 gives an RMS value of the difference in spatial information between the source and destination sequences.

Typical spatiotemporal metrics derived using the above parameters are briefly presented in Table. 4.1.

TABLE 4.1. Spatiotemporal Metrics

Spatiotemporal Metrics	**Type**	**Description**
Motion energy difference	FR, temporal	Added motion energy (error blocks, noise)
Repeated frames	FR, temporal	Lost motion energy (jerkiness)
Edge energy difference	FR, spatial	Dropped or repeated frames
Horizontal and vertical edges	FR, spatial	Added edge energy (edge noise, blockiness)
Spatial frequencies difference	FR, spatial	Lost edge energy (blur)

APPENDIX B. The Video Quality Experts Group (VQEG)

The VQEG was formed in 1997 to develop, validate, and standardize new objective measurement methods for video quality. The group is composed of experts from various backgrounds and organizations around the World. They are interested in FR/RR/NR quality assessment for various bandwidth videos for television and multimedia applications. VQEG has completed its Phase I test for FR video quality assessment for television in 2000 [37]. In Phase I test, 10 proposed video quality models (including several well-known models and PSNR)

were compared with the subjective evaluation results on a video database, which contains video sequences with a wide variety of distortion types and stimulus content. The result was, in some sense, surprising, since except for one or two proposed models that did not perform properly in the test, the other models performed statistically equivalent, including PSNR. Consequently, VQEG did not recommend any method for an ITU standard. VQEG is continuing its work on Phase II test for FR quality assessment for television, and RR/NR quality assessment for television and multimedia. Although it is hard to predict whether VQEG will be able to supply one or a few successful video quality assessment standards in the near future, the work of VQEG is important and unique from a research point of view. First, VQEG establishes large video databases with reliable subjective evaluation scores (the database used in the FR Phase I test is already available to the public), which will prove to be invaluable for future research on video quality assessment. Second, systematic approaches for comparing subjective and objective scores are being formalized. These approaches alone could become widely accepted standards in the research community. Third, by comparing state-of-the-art quality assessment models in different aspects, deeper understanding of the relative merits of different methods will be achieved, which will have a major impact on future improvement of the models. In addition, VQEG provides an ideal communication platform for the researchers who are working in the field.