VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

FACULTY OF ELECTRONICS

DEPARTMENT OF TELECOMMUNICATION ENGINEERING

Sara Blanco Fernández

MODELING QUALITY VIDEO METRICS OF VIDEO STREAMING OVER OPTICAL NETWORK

Final bachelor's work

Vilnius, 2009

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

FACULTY OF ELECTRONICS

DEPARTMENT OF TELECOMMUNICATION ENGINEERING

Sara Blanco Fernández

MODELING QUALITY VIDEO METRICS OF VIDEO STREAMING OVER OPTICAL NETWORK

Final bachelor‘s work

Head of department	Algumantas Kajackas
Supervisor	Dr. Šarūnas Paulikas
Supervisor UC3M	Dr. Carmen Vázquez
Academic Coordinator UC3M	Dr. Carmen Vázquez

Defended on 20th of February, 2009, Vilnius, Lithuania

Final result of evaluation: 8 or ECTS grade C

Tribunal: Prof. Dr. Habil. Algimantas Kajackas
Assoc. Prof. Dr. Sarunas Paulikas
Assoc. Prof. Dr. Arunas Saltis

Vilnius, 2009

CONTENTS

1. INTRODUCTION

Digital video data, stored in video databases and distributed through communication networks, is subject to various kinds of distortions during acquisition, compression, processing, transmission, and reproduction. Video quality is a characteristic of a video passed through a video transmission/processing system, a formal or informal measure of perceived video degradation (typically, compared to the original video) [1]. The impact of encoding and transmission impairments on the perceptual quality video streams is quite complex and depends heavily on the codec type and configuration, and on end system characteristics. Video processing systems may introduce some amounts of distortion or artefacts in the video signal.

According to the article [2] there are essentially three “models” for video performance measurement:

- **Full Reference** algorithms compare the output video stream to the input.
- **Zero Reference** algorithms analyze the output only.
- **Partial Reference** or **Reduced Reference** algorithms extract some parameters from the input stream and compare these to the equivalent parameters extracted from the output.

In fact, is not so easy to valuate image quality by comparing the output video with the input as some distortions could be clearly visible but not annoying to the observer. For example, if we take into consideration two images, one of them with another's luminance factor multiplied by a global factor, visual difference between them would be obviously but the observer would not appreciate lost of quality.

About another classification of measurement systems - according to several consulted literature [3]-[7] - it is possible to difference between two main groups: subjective quality metrics (using observers) and objective (mathematical metrics). We can difference within objective ones a new classification based on if we used characteristics of the HVS (Human Visual System) or not.

- **Error Measurement models** (mathematical metrics) get a quality metric without considering HVS. This would be an objective metric, related with the difference within the images and not with our perception.
- In a second group we include those models which incorporate more or less characteristics from the HVS. We call this second group **perceptual quality models.**
- Last group is **called hybrid quality measurement models**. Although they do not include HVS model they try to obtain values that fit the quality that observers feel by exploring new qualities related with perception.

Finally I am going to enumerate some measure models characteristics or properties that can tell us about goodness of the selected method [8]:

1. **Velocity:** is desirable that we obtain fast the value of the quality method. That is especially interesting when we want to use quality metrics to do better process or comprehension algorithms, quantification, etc.
2. **Cost:** computational cost depends on velocity and complexity of algorithms to get the results. We have to consider another additional cost, so to validate or evaluate the method it is necessary to take psychophysics test that have a temporally requisites and also number and observer characteristics requisites.
3. **Complexity:** It is strongly related with velocity and cost. The ideal would be to find the easiest method that could give a measure quality equal to that we perceive. In general, we get closed results to observatory results when we incorporate properties from HVS, which force to the method to be complex.
4. **Portability:** results that we get from the method shouldn't be altered if we repeat the same measurement within different times.
5. **Precision:** Refers how represents the value of the perception quality that the observer has.
6. **Robustness:** We pretend to have the valid results within a wide margin of parameters related to the measure (type of image, type of distortion, visibility conditions, etc.), That is to say we want robust methods.

7. **Type of result:** It can be numeric results (quality index) or maps of visual error. It depends on the application; it would be convenient one formula or the other one. In general the ideal method should have both responses.

Purpose of the Work

Package-switched communication networks, such as the Internet, can cause loss or severe delay of received data packages, depending on the network conditions and the quality of services. All these transmission errors may result in distortions in the received video data. I will investigate received video streaming over an optical network with heavy load and varying the network parameters, such as delay and jitter.

Video quality evaluation is an important problem to realize and quantify the video quality degradations that occur in the system, so that it can maintain, control and possibly enhance the quality of the video data. An effective image and video quality metric is crucial for this purpose. Subjective metrics are very difficult to carry out and we need an alternative mathematical measurement that we could use instead.

In the other hand, subjective metrics evaluation seems to be the most reliable way of assessing the quality of video, because human beings are the ultimate receivers in most applications. Furthermore some degradations may not be some important for observers than others. I will analyze several quality metrics to see which one fits better with subjective metrics.

First of all I will investigate different subjective and quality metrics, reviewing its weakness and strength points. Then I will implement a Matlab model to obtain results of the metrics and finally I will make various typical correlation models to investigate the relation between the different techniques to reach a conclusion about the goodness of the selected methods.

2. SUBJECTIVE IMAGE QUALITY MODELS

Subjective models are those which quality value is evaluated directly by an observer who is presented the images.

The most reliable way of assessing the quality of an image or video is subjective evaluation, because human beings are the ultimate receivers in most applications. The mean opinion score (MOS), which is a subjective quality measurement obtained from a number of human observers, has been regarded for many years as the most reliable form of quality measurement [6]. However, the MOS method is too inconvenient, slow and expensive for most applications.

One of the principle characteristics of these methods is that they give a more precise quality value at the moment due to be obtained directly from the observer. Otherwise, this kind of test is related to different inconvenient and we have to take into account that there would be a margin of error in the results. Next we will enumerate some problems about subjective methods [6]:

- It involves a high cost in time and a large number of people. To achieve a group of acceptable results it can take a few weeks. First of all, the place where the measure is going to take part should be equipped according to the recommendation, and we should take the test the maximum number of times as possible so we can achieve good results. The number of members and their capacities (age, profession, experience, etc.) can have an influence on the scores. To remove this dependency we have to choose a wide and heterogeneous group of observers.
- Recommendation ITU-r BT-500-10 presents one of the problems, which name is context effect. It is due to order and intensity of the distortions that appear. So, after several sequences or images with small distortions if there is an important distortion the observer is going to mark it with less score that would do normally.
- The set of values has also an influence on the goodness of the method. Discrete scales introduce an approximation which has to be compensated with a larger number of observers to reduce variance. Fixed scales favour observers not to use many extreme values.

- Another important inconvenient of these methods is that they are carried on under certain visibility conditions, and any small change on them would make the test invalid and it would be necessary to repeat the test.
- Subjective test do not provide spatial or temporal suitable information. The observer gives a global quality value but without taking into consideration where and when the error appears. These data are very useful to design codecs, watermarking methods, etc.

Source Signal for Audiovisual Tests

ITU recommendations suggest that the duration of the source sequences should be about 10 seconds for audiovisual, and the length of 5 short different sentences for speech. The termination of the sequences should not cause an incomplete sentence or musical phrase. An initial and a final silent period or gray scene, not longer than 500 ms, can be used to make the sequence be more natural. For the case of pair presentations, the references should have the best possible quality without any impairment. For audiovisual applications, speech and video should be perfectly synchronized.

Instructions for Assessors (Subjects)

Before carrying out the experiments, some instructions should be given to the assessors. These instructions include the method of assessment, the types of impairment or quality factors likely to occur, the grading scale, the sequence and timing. This information should be explained and given to the subjects in a written form. The range and type of impairments should be presented in preliminary trials. Training trials may be given to subjects to familiarize them with the task they will perform.

Number of Subjects and their Selection

As stated in ITU recommendations, the number of subjects required to carry out the subjective quality test can vary from 4 to 40. Four is the absolute minimum for statistical reasons. The number of assessors needed depends upon the sensitivity and reliability of the test procedure adopted. The average number of subjects is about 15. They should not be directly involved either in picture or audio quality evaluation as part of their work and should not be experienced assessors. Prior to a session, subjects should usually be screened for normal visual acuity or corrected-to-normal acuity and for normal colour vision (in the case of video quality evaluation).

The Test Sessions

Following the ITU recommendations, overall subjective tests should be divided into multiple sessions and each session should not last more than 30 minutes. For each session, we should add several dummy sequences (about four or five) at the beginning. These sequences should not be taken into account in the calculation. Their aim is to be used as training samples for the subjects to learn how to give meaningful rates. Furthermore, the reliability of subjects can be qualitatively evaluated by checking their behaviour when references pairs are given. In these cases, reliable subjects are expected to give evaluations very close to the maximum point in the quality scale.

Next we will enumerate the most common evaluating procedures.

2.1. Absolute Category Rating (ACR)

ACR is a category judgment method where the test sequences are presented one at a time and are rated independently on a category scale. This method is also called Single Stimulus Method. Subjects are asked to rate the quality of the presentation based on the level of the quality they have in their opinion for it after viewing or listening it. This phase is named the voting time. The voting time should be less than or equal to 10 seconds. The five-level scale for rating overall quality is the most used scale, see table 1. If higher discriminative power is required, a nine-level scale may be used, as shown in table 2. There is another variant of this scale which is the 11-point scale, depicted in table 3. Finally, there is a general scale, which is the continuous quality scale, see figure 1.

Table 1: ITU 5-point quality scale

Grading value	Estimated Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

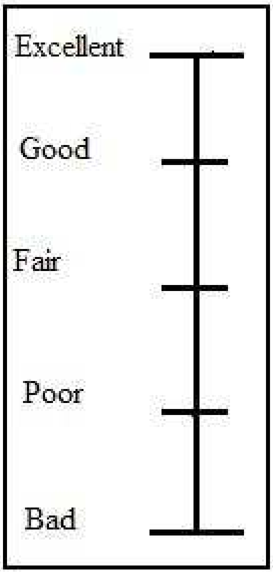
Table 2: ITU 9-point quality scale

Grading value	Estimated Quality
9	Excellent
8	
7	Good
6	
5	Fair
4	
3	Poor
2	
1	Bad

Table 3: Eleven-point quality scale - 10 score for the sequences that are identical to the reference one. 0 score is for the sequence that has no similarity with the reference.

Grading value	Estimated Quality
10	Best
9	Excellent
8	
7	Good
6	
5	Fair
4	
3	Poor
2	
1	Bad
0	Worst

Figure 1: Continuous scale



2.2. DSIS (Double Stimulus Impairment Scale)

This method is described in ITU-R BT.500-1. In this method videos are shown consequently in pairs: first one is the reference, and expert is informed about it, second one is impaired. After their playback, expert is asked to give his opinion using impairment scale: 5 imperceptible, 4 perceptible but not annoying, 3 slightly annoying, 2 annoying, 1 very annoying.

2.3. Degradation Category Rating (DCR)

In this second method, DCR, test sequences are presented in pairs. The first stimulus presented in each pair is always the source reference without any impairment. The second one is the same source but impaired by the test conditions. This method is also called the Double Stimulus Impairment Scale (DSIS) method. The voting time should be less than or equal to 10 sec. In this case the subjects are asked to rate the impairment of the second stimulus in relation to the reference. The five-level scale for rating the impairment is the most widely used one. However, all the quality scales used of ACR method can be used for DCR method but by replacing the quality adjectives by the corresponding impairment adjectives.

2.4. DSCQS (Double Stimulus Continuous Quality Scale) type I and type II

This method is described in ITU-R BT.500-11. In type I videos are played in pairs in one playback window. Each pair is repeated a given amount of times. During playback expert is free to switch between two videos. One of videos is the reference one, but expert is not informed about it. After playback expert is asked to give his opinion about each video sequence.

In type II (which is used more often) videos are played in pairs, and both videos are shown simultaneously. Each pair is repeated a given amount of times. As in type I, one of videos is the reference one, but expert is not informed about it. Impairment scale is the same as in type I.

Subjects are asked to assess the quality of both. The unimpaired one is included to serve as a reference, but the observers are not told which the reference sequence is. In the series of tests, the position of the reference is changed randomly. The subjects are asked to assess the overall sequence quality of each presentation by inserting a mark on a vertical scale. The vertical scales are printed in pairs to accommodate the double presentation of each test sequence. The scales are continuous to avoid quantizing errors, but they are divided into five equal lengths, which correspond to the normal ITU five-point quality scale. The associated terms categorizing the different levels are the same as those normally used.

2.5. SSCQE (Single Stimulus Continuous Quality Evaluation)

A continuous program is evaluated over long period (20-30 min.), scoring is a distribution of the amount of time. Reference is not shown. This method relates well to the time variant qualities of compressed television system. It allows viewers to dynamically rate the quality of an arbitrarily long video sequence using a slider mechanism with an associated quality scale. DSCQS scale is used.

An important issue in choosing a test method is the fundamental difference between methods that use explicit references (e.g. DCR or DSCQS) and methods that do not use any explicit reference (e.g. ACR). The latter does not test fidelity. The former, on the other hand, should be used when testing the fidelity of transmission with respect to the source signal. Thus, when it is important to check the fidelity with respect to the source signal, the DCR method should be used. Discrimination of imperceptible/perceptible impairment in the DCR scale supports this, as well as comparison with the reference quality. DSCQS, in addition, is used in the cases when the quality range is not completely covered. On the other hand, ACR is easy and fast to implement and the presentation of the stimuli is similar to that of the common use of the systems. Thus, ACR is well suited for qualification tests.

DSIS method is used to measure system robustness, for example, visible distortions caused by transmission errors. DSCQS evaluate a system in relation to a reference system and it works well for similar qualities due to its sensibility to small differences. SSCQE is used to, for example, fidelity measurements between two video distortional scenes.

Review of subjective methods and conclusions

The MOS methods have a lot of inconvenient, are slow and expensive for most applications. They also involve a high cost in time and a large number of people - to achieve a group of acceptable results it can take a few weeks.

The set of values has also an influence on the goodness of the method. Discrete scales introduce an approximation which has to be compensated with a larger number of observers to reduce variance. Fixed scales favour observers not to use many extreme values.

Despite results of these methods have a lot of disadvantages and carrying them out is difficult, they are necessary to prove precision of objective methods. There should be correlation between objective and subjective measurement results.

3. OBJECTIVE QUALITY IMAGE MODELS

They vary from a simple difference between original and distorted sequences to very complicated ones that are based on Human Vision System (HVS) models and include too complex mathematical calculations.

If we need to operate on both the original and the distorted video sequences, it is impossible to work in real time and to include these metrics in new design mechanisms. It is a limitation that can not allow rate control or video codecs to take into account the user's perception and the network factors.

A second disadvantage is that the obtained results are not always correlated with subjective data, thus they cannot measure correctly user's perception.

A third drawback is that they are very computationally extensive, especially the ones built based on VHS model, some of them cannot be used to evaluate the quality for video sequences of length greater than 1 sec.

Some of these metrics are designed and optimized basically to consider encoding impairments and restricted conditions, but they do not work efficiently when they are used in other conditions, for example distortion due to the transmission over the network.

One of the main advantages of subjective methods is that they provide us enough good immediately results and it is useful to monitor in a dynamic way image quality.

There are two different types of metrics, mathematical- that can be error simple objective measurements or can take into consideration perception of errors- and HVS based methods.

3.1. Error Measurement Models

We define error measurement models as those mathematical models which provide measurements based on simple mathematical functions, normally with spatial domain and point by point image process. These methods obtain a quality value in terms of deviations between processed images an original image, namely an error value [5].

Limitations

Use of this kind of metrics has a lot of advantages: simplicity, speed, and the most important, they have portability and they are globally accepted. The value obtained does not depend on neither specific metrics characteristics (monitor characteristics, viewing distance…) nor observer, which at first is an advantage. However, this is also a main problem due to if we don't take into consideration visual system sensitivity all the distortions are treated with the same importance, no matter which type, location or scene they are. They do not predict well the quality appreciated by the observer. In conclusion, these metrics are not very precise.

Another inconvenient is that these methods are not very robust to variations. In images with uniformed areas or specific content it is possible to make a good error prediction. They are useful, for example, to compare a group of images with the same reference image and the same distortion but with different value (different rate of compression, Gaussian noise, etc.)

However, when we used this metrics on different scene or in the same but with different kind of distortion, to equal PSNR values we can observe different qualities.

One last consideration can be about how to obtain results. PSNR, MSE or their variants give numerical results that can not capture spatial variations. This inconvenient can be overcome by using local sliding windows to obtain a spatial variation quality map.

Next I will define some quality metrics according to [10].

- **Mean Square Error (MSE)**

The simplest of image quality measurement is Mean Square Error (MSE). The large value of MSE means that image is poor quality. For a video sequence of frames each having x pixels with -bit depth, first the Mean Square Error (MSE) is calculated as follows:

$$MSE = \frac{1}{ME} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(x(m,n,k) - \hat{x}(m,n,k) \right)^2 \qquad (1)$$

where $x(m,n,k)$ and $\hat{x}(m,n,k)$ are the pixel luminance value in the *i, j* location in the *k* frame for the original and distorted sequences respectively.

The Root MSE (**RMSE**) is calculated using $RMSE = \sqrt{MSE}$

Besides, another quality metric comes from the MSE normalization to reduce sensitivity that appears in global changes in image intensity. The Normalized Mean Square Error (**NMSE**) normalizes image intensity, then standard deviation is one and average is zero. On these images we apply MSE formula.

When the two images are identical the MSE will be equal to zero.

- **Peak Signal to Noise Ratio (PSNR)**

The most commonly used objective quality metric is the Peak Signal to Noise Ratio (**PSNR**). The PSNR can be calculated as follows:

$$PSNR = 10 \log \frac{255^2}{MSE} \qquad (2)$$

MSE and RMSE measure the difference between the original and distorted sequences. PSNR measures the fidelity (how close a sequence is similar to an original one). Compared to other objective measures, PSNR is easy to compute and well understood by most researchers. The main problem is that the above measures consider only the luminance component, and neglect the chrominance one, which is important for human perception.

The small value of Peak Signal to Noise Ratio (PSNR) means that image is poor quality. Typical values for the PSNR in lossy image and video compression are between 30 and 50 dB, where higher is better. Acceptable values for wireless transmission quality loss are considered to be about 20 dB to 25 dB. A PSNR of zero can be obtained if the is completely white and K is completely black (or vice versa). When the two images are identical results in an infinite PSNR.

- **Structural Content (SC)**

The large value of Structural Content (SC) means that image is poor quality. SC is defined as follow:

$$SC = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} x(m,n)^2}{\sum_{m=1}^{M}\sum_{n=1}^{N} \hat{x}(m,n)^2} \quad (3)$$

- **Maximum Difference (MD)**

The large value of Maximum Difference (MD) means that image is poor quality. MD is defined as follow:

$$MD = Max\left(\left|x(m,n - \hat{x}(m,n)\right|\right) \quad (4)$$

- **Laplacian Mean Square Error (LMSE)**

This measure is based on the importance of edges measurement. The large value of Laplacian Mean Square Error (LMSE) means that image is poor quality. LMSE is defined as follow:

$$LMSE = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\left[L(x(m,n)) - L(\hat{x}(m,n))\right]^2}{\sum_{m=1}^{M}\sum_{n=1}^{N}\left[L(x(m,n))\right]^2} \quad (5)$$

- **Normalized Absolute Error (NAE)**

The large value of Normalized Absolute Error (NAE) means that image is poor quality. NAE is defined as follow:

$$NAE = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\left| x(m,n) - \hat{x}(m,n) \right|}{\sum_{m=1}^{M}\sum_{n=1}^{N}\left| x(m,n) \right|} \quad (6)$$

3.2. Perceptual Quality Measurement Models

These mathematical metrics not only take into consideration errors or distortion but also how visible they are in a particular image.

Quality measurement metrics provide numeric values that quantify the observer satisfaction with the processed image or visibility errors map.

The main advantage in this metric is that adding human visual systems characteristics can give as real quality results about how we perceive images. Comparatively these are complex computational models.

Limitations

Before description of most relevant models it is necessary to comment the weakest points of these methods. Perceptible quality, in all of them, is estimated by quantifying error visibility; we got it by introducing in the quality model the earliest steps function of HVS. However one of the main problems is the non-linearity and complexity of HVS.

We can not forget about lineal or almost lineal operators that we use for HVS modelling. It is globally accepted but simply, in comparison with its really complexity. This simplification makes that we make a sort of hypothesis in almost every model. First of all, we do not take into account high level processing like characteristics extraction, cognitive process, patron recognition and visual attention. It is assumed that quality perception is determined in the earliest steps of visual system. On the second place, we consider that CSF effect and masking in the same channel are main factors of perception. It is supposed that interaction between channels is enough small to ignore it. That is false, and there is a considerable effect between channels. Even though, most of models do not include masking between channels. Finally we do not take into account interaction between coefficients in the same channel after mask and CSF, considering them independent.

Even though with all these limitations, quality metric models have quite well results and are a good progress forward traditional metrics such as SNR and MSE.

- **The Structural Similarity Index**

SSIM index is a method for measuring the similarity between two images based on structural information [5]. It works under the assumption that human visual perception is highly adapted for extracting structural information from a scene. The SSIM index is a full reference metric, in other words, the measuring of image quality based on an initial uncompressed or distortion-free image as reference. Thus is based on the degradation of this structural information assuming that error visibility should not be equated with loss of quality as some distortions may be clearly visible but not so annoying. SSIM is designed to improve on traditional methods like PSNR and MSE, which have proved to be inconsistent with human eye perception. SSIM is also commonly used as a method of testing the quality of various lossy video compression methods.

First of all we make calculations on luminance, that is a mathematical mean, then we use contrast to calculate standard deviation and finally attributes that represent objects structure in the scene are modelled as a correlation.

The first component is structural comparision (correlation between original and distorted image), second component is luminance distortion and the third evaluates the difference in contrast values.

This result could be unstable if denominator sum is almost zero, and that for the method introduce some constants to avoid instability. Besides, statics are calculated locally, obtaining a local SSIM in each point of the image and using a sliding Gaussian window. We obtain a map of spatial quality and we got global SSIM by calculating a mean.

The SSIM metric is calculated on various windows of an image. The measure between two windows of the size NxN *x* and *y* is:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\,\mathrm{cov}_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (7)$$

with

- μ_x the <u>average</u> of *x* ;
- μ_y the <u>average</u> of *y* ;
- σ_x^2 the <u>variance</u> of *x* ;
- σ_y^2 the <u>variance</u> of *y* ;
- cov_{xy} the <u>covariance</u> of *y* ;
- $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ two variables to stabilize the division with weak denominator ;
- *L* the dynamic of the pixel-values ;
- k_1 = 0.01 and k_2 = 0.03 by default.

In order to evaluate the image quality this formula is applied only on the luminance. Typically it is calculated on window-sizes of 8x8. The window can be displaced pixel-by-pixel on the image but the authors propose to use only a subgroup of the possible windows to reduce the complexity of the calculation.

Finally SSIM does not attempt to predict image quality by accumulating the errors associated with psychophysically understood simple patterns, proposing to directly evaluate the structural changes between two complex-structured signals.

Between its advantages we can find we can remark its simplicity, portability, speed, and little computational cost. Besides, it provides not only a value but also a spatial quality map.

The SSIM index is a decimal value between 0 and 1. A value of 0 would mean zero correlation with the original image, and 1 means the exact same image. 0.95 SSIM, for example, would imply half as much variation from the original image as 0.90 SSIM. Through this index, image and video compression methods can be effectively compared.

- **Universal Quality Index (UQI)**

It is called Universal due to the quality measurement approach does not depend on the images being tested, the viewing conditions or the individual observers [20]. More importantly, it must be applicable to various image processing applications and provide meaningful comparison across different types of image distortions. Currently, the PSNR and MSE are still employed "universally", regardless of their questionable performance.

Let $x = \{x_i \mid i = 1,2,...N\}$ and $y = \{y_i \mid i = 1,2,...N\}$ be the original and the test image signals, respectively. The proposed quality index is defined as:

$$Q = \frac{4\sigma_{XY}\overline{x}\,\overline{y}}{\left(\sigma_X^2 + \sigma_Y^2\right)\left[\left(\overline{x}\right)^2 + \left(\overline{y}\right)^2\right]} \quad (8)$$

where

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (9) \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (10) \qquad \sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2 \quad (11)$$

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2 \quad (12) \qquad \sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) \quad (13)$$

The dynamic range of Q is [-1, 1]. The best value 1 is achieve if and only if $y_i = x_i$ for $y = 1,2,..,N$. The lowest value of -1 occurs when $y_i = 2\overline{x} - x_i$ for all $i = 1,2,..,N$. This quality index models any distortion as a combination of three different factors: loss of correlation, luminance distraction, and contrast distortion. In order to understand this, we rewrite the definition of Q as product of three components:

$$Q = \frac{4\sigma_{XY}}{\sigma_x \sigma_y} \cdot \frac{2\overline{x}\,\overline{y}}{\left(\overline{x}\right)^2 + \left(\overline{y}\right)^2} \cdot \frac{2\sigma_x \sigma_y}{\sigma_X^2 + \sigma_Y^2} \quad (14)$$

The first component is the correlation coefficient between x and y, which measures the degree of linear correlation between x and y, and its dynamic range is [-1, 1]. The best value 1 is obtained when $y_i = ax_i + b$ for all $i = 1,2,..,N$, where *a* and *b* are constants and a>0. Even if x and y are linearly related, there still might be relative distortions between them, which are evaluated in the second and third components. The second component, with a value range of [0,1], measures how close the mean luminance is between x and y. It equals 1 if and only if $\overline{x} = \overline{y}$. σ_x and σ_y can be viewed as estimate of the contrast of x and y, so the third component measures how similar the contrasts of the images are. Its range of values is also [0,1], where the best value 1 is achieved if and only if $\sigma_x = \sigma_y$.

Image signals are generally non-stationary while image quality is often space variant. It is possible to apply this quality measurement method to local regions using a sliding window approach to measure statistical features locally and then combine them together so we can evaluate an entire image using a single overall quality value.

3.3. Hybrid quality metrics models

In this group we include models which try to obtain results which fits with the perceptible measurement, but without including HVS.

- **VQM model**

These automated measurement algorithms provide close approximations to the overall quality impressions, or mean opinion scores, of digital video impairments that have been graded by panels of viewers.

VQM is a hybrid model that does not include HVS model but explore qualities related with perception. The VQM consists of a linear combination of four parameters that have been optimized for the standard viewing distance of six times picture height. Three parameters are extracted from spatial gradients of the luminance component (Y) input and output video streams while one parameter is extracted from the vector formed by the chrominance components (CB, CR).

In addition to providing technology-independent perception-based estimates of subjective quality, the VQM has low computational complexity and can be used for continuous real-time in-service quality monitoring applications.

The NTIA General VQM has been shown to be highly correlated to subjective ratings of processed video clips from an HDTV experiment that included a fairly wide range of codecs, bit rates, and even some transmission errors. The Pearson linear correlation coefficient between VQM and each of the individual subjective data sets achieved an average coefficient of 0.90.

ITU-T J.144 does not actually specify a single algorithm but provides guidelines on the selection of appropriate techniques. J.144 does contain descriptions and test results for four full reference algorithms.

In the ITU-R BT.500-10 recommendation "Methodology for the subjective assessment of the quality of television pictures" are regulated some of the tests for static image and video subjective evaluation. The norm includes how to choose test materials and observers, visibility conditions, evaluation procedures and data analysis.

The NTIA General VQM scores are reported on a nominal range of [0, 1], where zero indicates excellent quality.

Review of subjective methods and conclusions

As I commented before, the use of this kind of metrics has a lot of advantages: simplicity, speed, and the most important, they have portability and they are globally accepted. The value obtained does not depend on neither specific metrics characteristics (monitor characteristics, viewing distance…) nor observer, which at first is an advantage. However, this is also a main problem due to if we don't take into consideration visual system sensitivity all the distortions are treated with the same importance, no matter which type, location or scene they are.

I will use the above mathematical formulas (1)-(8) in matlab for a subsequent analysis. I will obtain quality measures with the formulas component by component and frame by frame and then I calculate the mean for each video sample.

An important observation is that not all the quality-affecting parameters can be considered. For example, the frame rate effect cannot be considered if we compare the original and distorted. This means that both sequences must have the same frame rate and the decoded picture of the processed sequence must correspond to the encoded picture in each frame of the original sequence, otherwise the results will degrade.

4. IMPLEMENTATION OF QUALITY MODELS

I have received 38 samples that correspond to two different videos. These videos have been compressed to MPEG-2 and then a simulation was made about transmitting them over an optical network with heavy load and varying different parameters of the network. According to literature [10] and as the videos were compressed and then added some impairment I expect some video degradation that I am going to enumerate next.

Problems due to transmission impairments

1. Lost packets lead to missing blocks within the decoded image - causing "blockiness" and if a large proportion of blocks are missing - frame freeze. The impact of a lost packet will vary considerably, depending on the type of frame impacted. If an I or P frame is corrupted then the resulting image degradation will affect all the following P and B frames until the next I frame is received. As a P frame generally represents a smaller region of the image than an I frame, the effect of packet loss on P frames will be slightly less than on an I frame. B frames are not used as reference frames and hence a lost B frame packet will only affect that frame.

Problems due to encoder and compression

1. Block distortion: can be caused by coarse quantization of the spatial frequency components of an image during encoding, and is due to the block structure of MPEG images.
2. Blurring: Blurring is a reduction in the sharpness of edges, and will be more widely observed in lower bit rate or lower frame rate algorithms or on video sequences with high rates of motion.
3. Edge Busyness: is caused by quantization of the image at the boundaries between areas with a significantly different colour or brightness level.

4. Mosquito Noise: is a form of edge busyness distortion that is associated with movement within the image which results in moving artefacts or noise patterns superimposed over the moving object.
5. Quantization Noise: typically occurs as visible “noise” (snow) over most of the image and will not necessarily be uniform.
6. Jerkiness: Is typically associated with low bit rate encoding of video sequences with motion. Motion that was originally smooth appears as discontinuous “jumps”.
7. Color Pixellation: Blocks of coloured pixels are typically due to errors that occur during the transcoding process when a digital image is converted from one format to another.

Then I will implement a Matlab model to obtain results of various metrics in order to make a future analysis of which can be the best metric for these special conditions.

I have implemented the following subjective quality measurements: SSIM, UQI, MSE, PSNR, SC, MD, LMSE and NAE using streamingQuality.m, ssim.m, uqi.m and iq_measures.m. A matlab version R2007b is required in order to use `mmreader()` function.

I obtain quality measures component by component and frame by frame and then I calculate the mean for each video sample.

I have used *foreman.avi* and *hall_monitor_qcif* video samples obtained of varying jitter and delay to execute streamingQuality.m in order to obtain our quality results as the following example:

```
>>streamingQuality('foreman_original,avi','foreman45 _processed.avi')
```

Then I have used the external program ‘bvqm_pc_v12’ to obtain VQM quality for each video sample and save the results in the following .mat fies: SSIM.mat, UQI.mat, MSE.mat, PSNR.mat, SC.mat, MD.mat, LMSE.mat, NAE.mat and VQM.mat, where 18 first data correspond to foreman.avi and the last 18 data correspond to hall_monitor_qcif.

The results obtained for each video sample varying delay and jitter parameters are detailed in table 4.

Table 4: Quality results for each video

VIDEO SAMPLE	SSIM	UQI	MSE	PSNR	SC	MD	LMSE	NAE	VQM
foreman_d=31,4 j=5	0.9718	0.9711	71.1259	30.3233	1.035	71.4938	0.0407	0.0475	0.1901
foreman_d=31,4 j=6	0.9728	0.972	68.8055	30.373	1.0349	70.8979	0.0395	0.0471	0.1881
foreman_d=31,4 j=7	0.9713	0.9705	74.7692	30.2276	1.035	72.2357	0.0415	0.048	0.1946
foreman_d=31,5 j=5	0.9749	0.9741	64.7947	30.5122	1.0345	69.2738	0.0371	0.046	0.1849
foreman_d=31,5 j=6	0.974	0.9732	68.1649	30.3931	1.0347	70.3468	0.039	0.0467	0.192
foreman_d=31,5 j=7	0.9651	0.9642	92.7715	29.8609	1.0346	77.0606	0.0519	0.0504	0.2183
foreman_d=31,6 j=5	0.9701	0.9693	81.4847	30.182	1.0351	72.7273	0.0441	0.0486	0.2132
foreman_d=31,6 j=6	0.9671	0.9662	87.8021	29.9979	1.0347	74.5859	0.0505	0.0499	0.2192
foreman_d=31,6 j=7	0.9581	0.9574	158.8296	29.393	1.0439	81.5533	0.0642	0.0565	0.229
foreman_d=31,7 j=5	0.9729	0.972	77.9616	30.3339	1.0348	71.0673	0.0419	0.0473	0.1929
foreman_d=31,7 j=6	0.9601	0.959	191.197	29.5751	1.453	78.5264	0.0625	0.0563	0.2378
foreman_d=31,7 j=7	0.9615	0.9604	139.128	29.7173	1.0395	77.0853	0.0614	0.0532	0.2367
foreman_d=31,8 j=5	0.9705	0.9697	88.4153	30.2874	1.0361	71.6117	0.0473	0.0479	0.2013
foreman_d=31,8 j=6	0.9541	0.9532	160.2086	29.194	1.0406	83.3547	0.0725	0.057	0.2621
foreman_d=31,8 j=7	0.9576	0.9566	130.5326	29.4284	1.0368	80.3154	0.0668	0.0541	0.2562
foreman_d=31,9 j=5	0.9686	0.9678	84.8308	30.0399	1.0344	74.7486	0.0468	0.0494	0.2151
foreman_d=31,9 j=6	0.9593	0.9583	167.7229	29.6434	1.453	77.8384	0.0622	0.0548	0.2308
foreman_d=31,9 j=7	0.9398	0.9386	242.2549	28.3058	1.0453	91.9315	0.1005	0.0647	0.2772
hall_monitor_qcif_d=31,4 j=5	0.9793	0.9773	50.8856	31.2954	1.0135	83.1089	0.0441	0.0384	0.1007
hall_monitor_qcif_d=31,4 j=6	0.9792	0.9772	51.0907	31.2874	1.0135	82.7677	0.0442	0.0384	0.1018
fhall_monitor_qcif_d=31,4 j=7	0.9784	0.9763	53.1657	31.1876	1.0135	84.7239	0.046	0.0389	0.1055
hall_monitor_qcif_d=31,5 j=5	0.9789	0.9769	51.8248	31.2463	1.0134	83.7351	0.045	0.0386	0.104
hall_monitor_qcif_d=31,5 j=6	0.9793	0.9773	50.9833	31.2959	1.0135	82.8462	0.0441	0.0384	0.1008
hall_monitor_qcif_d=31,5 j=7	0.9783	0.9762	53.3943	31.1494	1.0134	86.1695	0.0464	0.0391	0.1085
hall_monitor_qcif_d=31,6 j=5	0.9793	0.9773	50.688	31.2941	1.0135	83.0539	0.044	0.0384	0.996
hall_monitor_qcif_d=31,6 j=6	0.9786	0.9766	52.8871	31.2007	1.0135	84.7924	0.0457	0.0388	0.1027
hall_monitor_qcif_d=31,6 j=7	0.9777	0.9755	54.9348	31.0612	1.0134	88.0269	0.0478	0.0394	0.1115
hall_monitor_qcif_d=31,7 j=5	0.9789	0.9769	51.8946	31.2539	1.0135	83.4265	0.0449	0.0386	0.1033
hall_monitor_qcif_d=31,7 j=6	0.9782	0.976	53.5818	31.1405	1.0135	86.5331	0.0465	0.0391	0.11
hall_monitor_qcif_d=31,7 j=7	0.9763	0.9741	100.505	31.0504	1.453	87.3906	0.0494	0.0419	0.1134
hall_monitor_qcif_d=31,8 j=5	0.9787	0.9766	52.4181	31.2221	1.0135	84.5488	0.0454	0.0387	0.1096
hall_monitor_qcif_d=31,8 j=6	0.9784	0.9763	52.9137	31.181	1.0134	85.8081	0.046	0.0389	0.107
hall_monitor_qcif_d=31,8 j=7	0.9772	0.9749	56.2779	30.9991	1.0135	89.5275	0.0489	0.0397	0.1167
hall_monitor_qcif_d=31,9 j=5	0.9792	0.9771	51.0913	31.2818	1.0135	82.7262	0.0442	0.0386	0.1011
hall_monitor_qcif_d=31,9 j=6	0.977	0.9748	59.4246	31.0313	1.0143	88.5533	0.0487	0.0401	0.1137
hall_monitor_qcif_d=31,9 j=7	0.9754	0.9732	81.6317	30.8975	1.453	91.477	0.0518	0.0419	0.1191

5. ANALISYS OF CORRELATION BETWEEN QUALITY ESTIMATATIONS OF DIFFERENTS ALGORITMS

The performance of an objective video quality metric is evaluated by computing the correlation between the objective scores and the subjective test results. The last ones are usually the ones called mean opinion score (MOS).

To analyze correlation between quality algorithms and between each one and results closed to subjective ones we will use VQM instead of MOS, due to this is an automated algorithm that can provide closed approximation to MOS-Pearson correlation of 0.9-. In this way I would not need to carry out the subjective tests and avoid all MOS drawbacks.

Next I will represent the results in the following figures in order to have a visual idea of what I can obtain in the following analysis. The scales go from good to bad or vice versa depending on the method scale. I will explain the normal value results of each method:

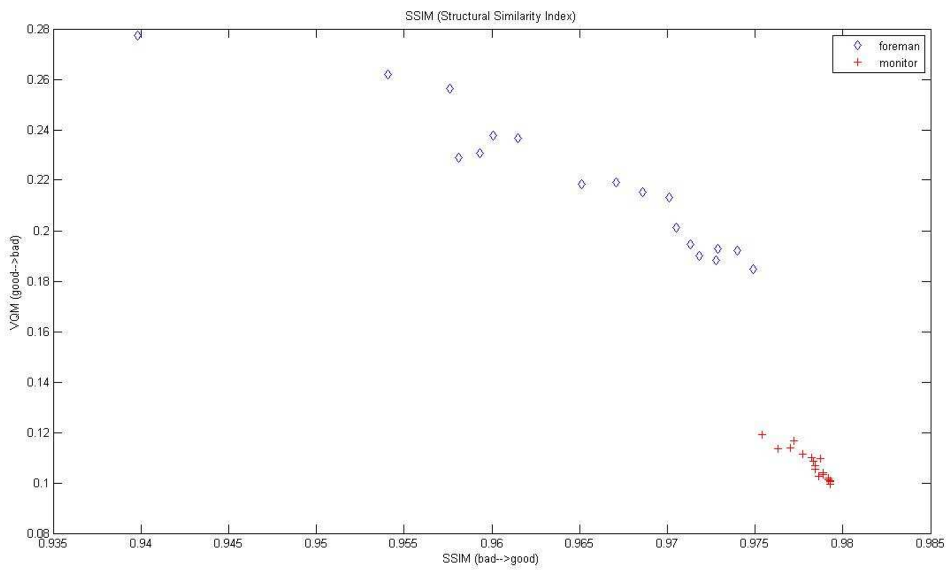


Figure 2: SSIM-VQM

The SSIM index is a decimal value between 0 and 1. A value of 0 would mean zero correlation with the original image, and 1 means the exact same image. 0.95 SSIM, for example, would imply half as much variation from the original image as 0.90 SSIM. SSIM have a negative correlation with VQM.

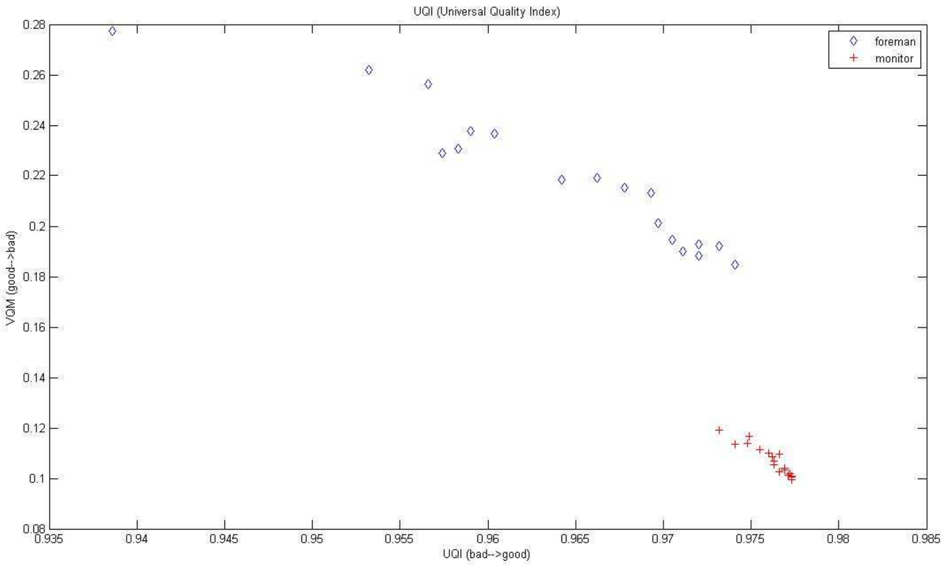


Figure 3: UQI-VQM

UQI has a negative correlation with VQM. Its range of values is also [0,1], where the best value 1 is achieved if both video sample are identical.

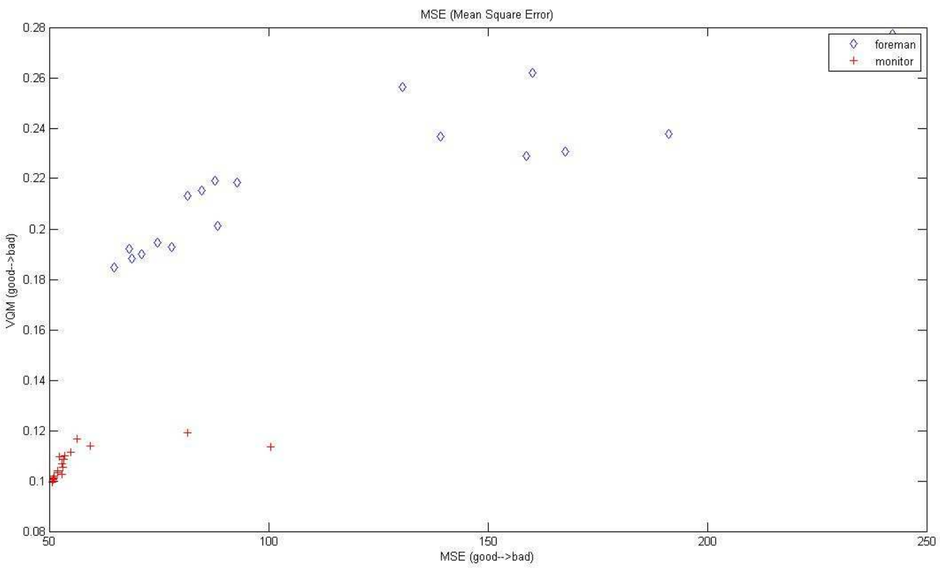


Figure 4: MSE-VQM

The large value of MSE means that image is poor quality. This method has a positive correlation with VQM.

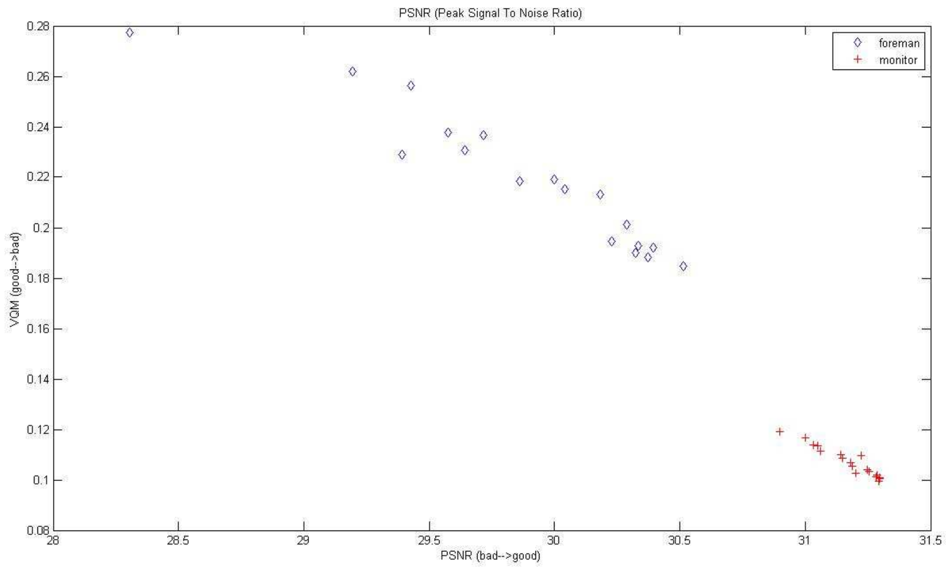


Figure 5: PSNR-VQM

The small value of Peak Signal to Noise Ratio (PSNR) means that image is poor quality. Typical values for the PSNR in lossy image and video compression are between 30 and 50 dB, where higher is better. Acceptable values for wireless transmission quality loss are considered to be about 20 dB to 25 dB. A PSNR of zero can be obtained if the I is completely white and K is completely black (or vice versa). When the two images are identical it results in an infinite PSNR. PSNR has a negative correlation.

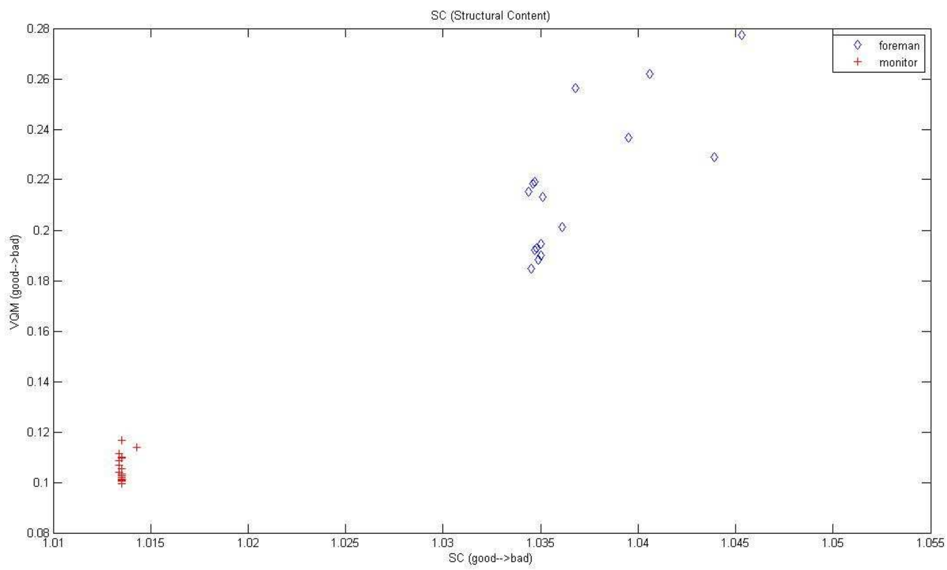


Figure 6: SC-VQM

The large value of Structural Content (SC) means that image is poor quality. SC has a positive correlation.

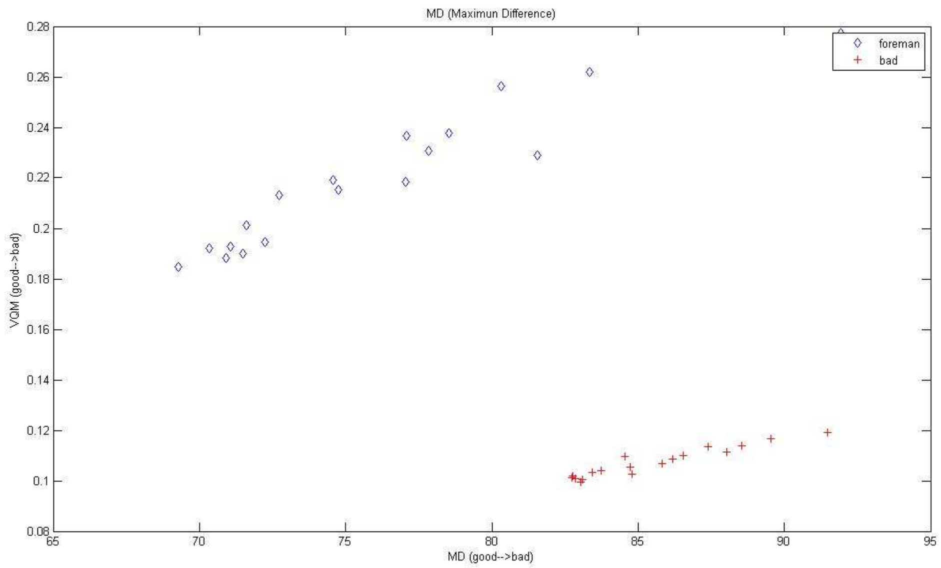


Figure 7: MD-VQM

The large value of Maximum Difference (MD) means that image is poor quality. It has a positive correlation with VQM.

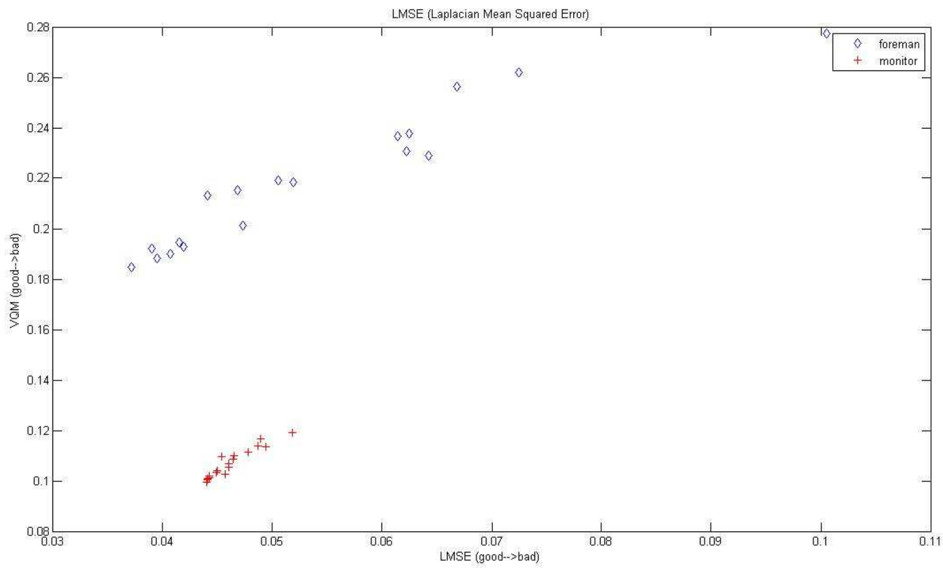


Figure 8: LMSE-VQM

The large value of Laplacian Mean Square Error (LMSE) means that image is poor quality. LMSE has a positive correlation with VQM.

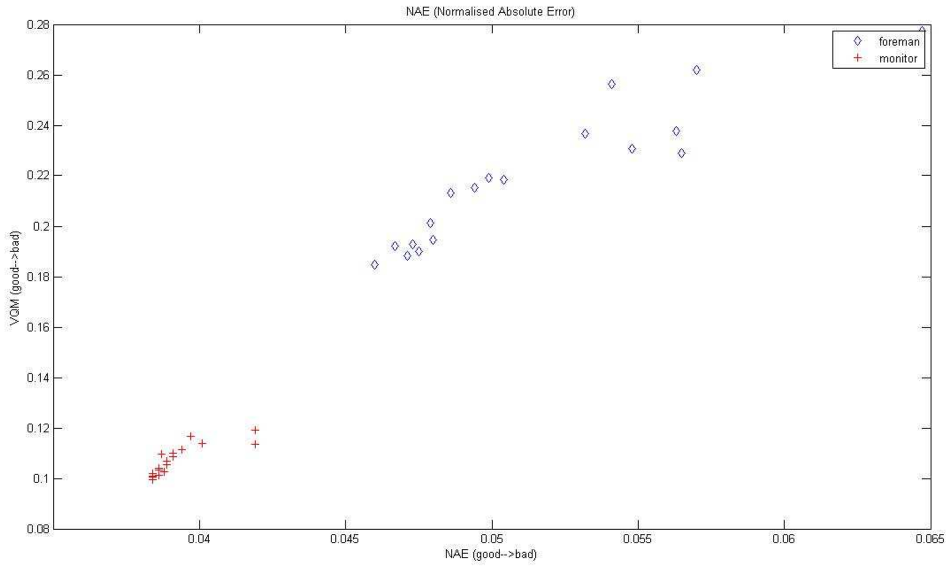


Figure 9: NAE-VQM

The large value of Normalized Absolute Error (NAE) means that image is poor quality. It has a positive correlation with VQM.

From these figures we can realize that there is a noticeable difference between 'foreman.avi' video samples and hall_monitor_qcif video samples. Monitor samples always have better results in every metric. This could be due to the fact that these samples have more static frames and packet loss does not affect so much as in a more dynamic video like foreman samples.

The relations between objective measurement and VQM measurement of the video compressed samples shown in plots 8-4 respectively seems to be closer to a linear relation than the others.

In the following sections I will explain the most frequently used correlation coefficients to analyze video quality, these are: linear correlation coefficient, Spearman's rank correlation coefficient, kurtosis and Kappa coefficient. I will implement all of them in order to get some conclusions.

5.1. Linear Correlation Coefficient

Definition

The quantity *r*, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient, in honour of its developer Karl Pearson.

The mathematical formula for computing *r* is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} \quad (15)$$

where n is the number of pairs of data.

Remarks

- ✓ The value of *r* is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.
- ✓ Positive correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.
- ✓ Negative correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
- ✓ No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables. Note that r is a dimensionless quantity; that is, it does not depend on the units employed.
- ✓ A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative.

✓ A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the type of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Description of the analysis

```
>> LinearCorrelation
```

- I load .mat files with quality results for each quality algorithm for different network parameters and different video samples.
- Then I obtain a matrix that contains all results.

```
>> d=[UQI MSE PSNR SC MD LMSE NAE VQM];
>>d=d';
```

- Transpose the matrix and apply `corrcoef()` matlab function to obtain a correlation matrix: R (diagonal 1 and values between -1 and +1).

```
>>R=corrcoef(d);
```

Results and Conclusions

R matrix:

Table 5: Linear Correlation

	SSIM	UQI	MSE	PSNR	SC	MD	LMSE	NAE	VQM
SSIM	1	0.9985	-0.9519	0.9749	-0.7530	0.1037	-0.8821	-0.9667	-0.9002
UQI	0.9985	1	-0.9574	0.9622	-0.7317	0.0507	-09060	-0.9525	-0.8765
MSE	-0.9519	-0.9574	1	0.9000	-0.7573	0.0063	0.8944	0.9149	0.8033
PSNR	0.9749	0.9622	0.9000	1	-0.8240	0.2942	-0.7619	-0.9947	-0.9692
SC	-0.7530	-0.7317	0.7573	0.8240	1	-0.4188	0.4813	0.8551	0.8445
MD	0.1037	0.00507	0.0063	0.2942	-0.4188	1	0.3586	-0.3270	-0.4969
LMSE	-0.8821	-0.9060	0.8944	-0.7619	0.4813	0.3586	1	0.7412	0.5987
NAE	-0.9667	-0.9525	-0.9947	0.8551	0.8551	-0.3270	0.7412	1	0.9702
VQM	-0.9002	0.8033	-0.9692	0.8445	0.8445	-0.4969	0.5987	0.9702	1

✓ As we can see in the table 5 the most related measurements with VQM metric are <u>NAE</u>, <u>PSNR</u>, and <u>SSIM</u>. UQI, MSE and SC have also strong correlation whereas MD and LMSE have a weak one. In deed these are suspected results if we observe the figures 2-9.

✓ We also notice that MD has weak correlation with the other metrics.

5.2. Spearman's Rank Correlation Coefficient

Definition

Spearman's rho is a measure of the linear relationship between two variables. It differs from Pearson's correlation only in that the computations are done after the numbers are converted to ranks. The Spearman Rank Order Correlation Coefficient was developed by Spearman to use with this type of data. When converting to ranks, the smallest value on a group of data becomes a rank of 1, etc. The Symbol is r_s. Is a non-parametric measure of correlation – that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables.

The mathematical formula for computing r_s is:

$$r_s = 1 - \frac{6\sum d^2}{N3 - N} \quad (16)$$

where 6 is a constant (it is always used in the formula), D refers to the difference between a subjects ranks on the two variables and N is the number of subjects.

Remarks

- ✓ As in linear correlation coefficient the value of r_s is such that $-1 \leq r_s \leq +1$. Zero means no correlations between variables while ± 1 is perfect correlation.

Description of the analysis

- Convert to ranks the results for each of the metrics.

```
>> ranking
```

- Then I save results in .mat files: SSIMrank.mat, UQIrank.mat, MSErank.mat, PSNRrank.mat, SCrank.mat,MDrank.mat, LMSErank.mat, NAErank.mat and VQMrank.mat.

- Note: Those metrics which have negative correlation -as we can see in table 5- with VQM have and inverse rank -SSIM, UQI, PSNR, MD-.
- We use the above data when execute spear.m.

```
>> spear
```

- Obtain the difference between ranks of each measurement and VQM rank, and power two of the difference.
- Summatory of all the differences powered two.
- Apply the equation.
- I obtain the spearman coefficient.

Results and Conclusions

Table 6: Spearman Correlation

	SSIM	UQI	MSE	PSNR	SC	MD	LMSE	NAE
VQM	0.9068	0.8921	0.4997	0.9905	0.8100	0.5667	0.7801	0.9821

- ✓ With Spearman correlation I get similar results than with Pearson.
- ✓ In relation with Spearman coefficients the best results are obtained again by NAE, PSNR and SSIM.
- ✓ UQI and SC have also strong correlation with VQM and the weakest relations are VQM-MSE and VQM-MD.

5.3. Kappa Coefficient

Definition

Kappa provides a measure of the degree to which two judges, A and B, concur in their respective sorting of N items into k mutually exclusive categories. A 'judge' in this context can be an individual human being, a set of individuals who sort the N items collectively, or some non-human agency, such as a computer program or diagnostic test, that performs a sorting on the basis of specified criteria.

The kappa coefficient is defined as:

$$k = \frac{(p_o - p_e)}{(1 - E_e)} \quad (17)$$

where *pe* is the observer percentage agreement and *po* is the probability of random agreement.

Remarks

✓ The original and simplest version of kappa is the unweighted kappa coefficient introduced by J. Cohen in 1960. The measures of weighted kappa are meaningful only if the categories are ordinal and if the weightings ascribed to the categories faithfully reflect the reality of the situation. The weightings in this case are determined by the imputed relative distances between successive ordinal categories.

Description of the analysis

- We need a square data matrix of VQM and each objective measurement results for each of the network parameters.

```
>>kappaMatrix
```

- Run `kappa.m` to calculate the observed percentage agreement is *po* and the *pe*, the probability of random agreement.

```
>>kappa
```

- Analyze agreement between VQM and each of the metrics.
- Decide if we accept or reject the hypothesis:
    - Null (Ho) -There is no association between the variables.
    - Alternate (Ha): There is an association between the variables.

Table 7: SSIM Kappa

-------SSIM--------	
Observed agreement (po)	0.5378
Random agreement (pe)	0.3549
Agreement due to true concordance (po-pe)	0.1728
Residual not random agreement (1-pe)	0.6451
Cohen's kappa	0.2679
Fair agreement	
Ha→Reject null hypotesis: observed agreement is not accidental	

Table 8: UQI Kappa

-------UQI--------	
Observed agreement (po)	0.5278
Random agreement (pe)	0.3827
Agreement due to true concordance (po-pe)	0.1451
Residual not random agreement (1-pe)	0.6173
Cohen's kappa	0.2350
Fair agreement	
Ho→Accept null hypotesis: observed agreement is accidental	

Table 9: MSE Kappa

-------MSE-------	
Observed agreement (po)	0.4444
Random agreement (pe)	0.4028
Agreement due to true concordance (po-pe)	0.0417
Residual not random agreement (1-pe)	0.5972
Cohen's kappa	0.0698
Slight agreement	
Ho→Accept null hypotesis: observed agreement is accidental	

Table 10: PSNR Kappa

-------PSNR-------	
Observed agreement (po)	0.5556
Random agreement (pe)	0.3032
Agreement due to true concordance (po-pe)	0.2523
Residual not random agreement (1-pe)	0.6968
Cohen's kappa	0.3621
Fair agreement	
Ha→Reject null hypotesis: observed agreement is not accidental	

Table 11: SC Kappa

-------SC-------	
Observed agreement (po)	0.6111
Random agreement (pe)	0.3148
Agreement due to true concordance (po-pe)	0.2963
Residual not random agreement (1-pe)	0.6852
Cohen's kappa	0.4324
Moderate agreement	
Ha→Reject null hypotesis: observed agreement is not accidental	

Table 12: MD Kappa

------MD------	
Observed agreement (po)	0.2500
Random agreement (pe)	0.1512
Agreement due to true concordance (po-pe)	0.0988
Residual not random agreement (1-pe)	0.8488
Cohen's kappa	0.1164
Slight agreement	
Ho→Accept null hypotesis: observed agreement is accidental	

Table 13: LMSE Kappa

-------LMSE-------	
Observed agreement (po)	0.5000
Random agreement (pe)	0.3850
Agreement due to true concordance (po-pe)	0.1150
Residual not random agreement (1-pe)	0.6150
Cohen's kappa	0.1870
Slight agreement	
Ho→Accept null hypotesis: observed agreement is accidental	

Table 14: NAE Kappa

-------NAE-------	
Observed agreement (po)	1.0000
Random agreement (pe)	0.3441
Agreement due to true concordance (po-pe)	0.6559
Residual not random agreement (1-pe)	0.69559
Cohen's kappa	1.0000
Perfect agreement	
Ho→Reject null hypotesis: observed agreement is not accidental	

- ✓ Best results are obtain by NAE (perfect agreement), SC (moderate agreement), SSIM (fair), PSNR (fair).
- ✓ Worst results are obtained by MD, LMSE and MSE (slight agreement).

5.4. Kurtosis

Definition

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

The kurtosis of a distribution is defined as:

$$k = \frac{E(x-\mu)^4}{\sigma^4} \quad (18)$$

where μ is the mean of x, σ is the standard deviation of x, and $E(t)$ represents the expected value of the quantity t.

Remarks

- ✓ The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3.
- ✓ Some definitions of kurtosis subtract 3 from the computed value, so that the normal distribution has kurtosis of 0. The kurtosis function in matlab k=kurtosis(X) does not use this convention.

Description of the analysis

- I use Matlab function kurtosis, `k = kurtosis(X)` returns the sample kurtosis of X. For vectors, kurtosis(X) is the kurtosis of the elements in the vector X. For matrices kurtosis(X) returns the sample kurtosis for each column of X.

- Then I compare the kurtosis of each quality measurement results for all videos with VQM kurtosis.

- If the metrics have a high value could mean that is a sensitive method for the network parameters.

Results and Conclusions

Table 15: Kurtosis with all samples

VQM	SSIM	UQI	MSE	PSNR	SC	MD	LMSE	NAE
1.5199	5.1622	5.7247	5.233	3.0817	1.2456	1.9116	10.0953	2.7091

Table 16: Kurtosis with foreman samples

VQM	SSIM	UQI	MSE	PSNR	SC	MD	LMSE	NAE
2.3538	4.2341	4.2462	3.0629	4.4049	2.1059	4.2457	4.8302	3.6745

Table 17: Kurtosis with monitor samples

VQM	SSIM	UQI	MSE	PSNR	SC	MD	LMSE	NAE
2.0796	3.5426	3.2344	8.6723	2.5959	7.1195	2.5787	3.1633	4.8371

- ✓ SSIM and UQI have very similar results as in the figures 2-3.
- ✓ LMSE give similar results for both type of video samples. Also UQI and SSIM.
- ✓ SC and MSE do not recognize as many losses as the other metrics in statics video sample but have fair better results with foreman samples. In the other hand, PSNR, MD and VQM obtain very good results.

6. CONCLUSIONS

Conclusions of this studio about the effect of jitter on the perceived quality were:

- The effects of jitter on the degradation of the perceived quality are very similar to the effects of the losses. Moreover, the degradation is severe in the presence of low levels of loss and/or jitter, while higher values of these parameters do not degrade the perceived quality proportionately.
- As I expected, those statics sequences with less temporal variation- were more robust against the presence of jitter than those ones with higher temporal variation.

Conclusions of this study about quality estimation algorithms:

- Robustness of the metric is an important parameter; we need to have valid results within a wide margin of parameters related to the measure - type of video and type of distortion. SSIM and UQI seams to be the most robust methods, it should be due to they not only take into consideration errors or distortion but also how visible they are in a particular image.
- The highest reliability of objective quality measurements for MPEG-2 compressed videos in this studio are obtained with NAE and PSNR. Then, these are the best measurement and suitable for streaming video quality evaluation and are the metrics that I recommend for this purpose. The relations between objective measurement and VQM measurement of the video compressed samples shown in figures 9-5 respectively seems to be closer to a linear relation than the others.
- The less correlated metrics with VQM where LMSE, MSE and MD. As can be seen, the scatter plots of LMSE, MSE and MD are more widely clustered.

7. SUMMARY

1. Problemas del Video Streaming

Video streaming es video digital transmitido por redes de comunicación y reproducido inmediatamente tras la recepción. Debido a la transmisión sobre redes, el video esta sometido a pérdidas de paquetes, delay y jitter. También es necesario comprimir el video previo a la transmisión de este.

Como resultado de la compresión, el procesamiento, la transmisión y reproducción, el video puede presentar diferentes tipos de distorsión que se enumeran a continuación:

1. Blockiness: Efecto en imágenes normalmente debido a simplificación en compresión. Normalmente en compresión JPEG la imagen es dividida en bloques de 8x8, sobre los cuales se realiza algún tipo de simplificación de colores, para disminuir la cantidad de información a procesar, de manera independiente del resto de la imagen. Cuando se utiliza compresión muy alta, los bordes de los bloques se hacen demasiado evidentes al descomprimir, produciendo este efecto.
2. Blurring: Es una reducción en la forma de los bordes en la imagen, esta distorsión se observará mejor con algoritmos con bajo bit rate o frame rate o en secuencias de video con mucho movimiento.
3. Edge Busyness: causada por la cuantificación de la imagen en las fronteras entre diferentes áreas con una diferencia significativa entre colores o niveles de brillo.
4. Mosquito Noise: es una forma de distorsión edge busyness asociada con el movimiento de la imagen. El resultado es artefactos en movimiento o patrones de ruido superpuestos sobre un objeto en movimiento.
5. Quantization Noise: Ocurre típicamente como ruido visible “nieve” sobre la mayoría de la imagen y no necesariamente de forma uniforme.

6. Jerkiness: Esta típicamente asociado con tasas bajas de codificación de bit de secuencias de video con movimiento. Movimiento que originalmente era suave aparece como saltos discontinuos.
7. Color Pixellation: Bloques de pixeles coloreados debido a errores que ocurren durante el proceso de transcodificación cuando una imagen digital es convertida de un formato a otro.

A continuación se van a enumerar algunas de las características o propiedades de los modelos de calidad sobre las que se puede evaluar la bondad del método seleccionado. Entre ellas se pueden destacar las siguientes:

1. Velocidad: es deseable que el valor de calidad que resulta de la utilización del método se obtenga de forma rápida. Esto adquiere un especial interés cuando se va a hacer uso de la medida de calidad para mejorar los procesos o algoritmos de compresión, cuantificación...
2. Coste: el computacional depende de la velocidad y de la complejidad de los algoritmos para la obtención de un resultado. Además, se deben considerar otros costes adicionales ya que para la validación o evaluación del método es necesario llevar a cabo test psicofísicos que tienen una serie de requisitos temporales y también en cuanto a número y características de los observadores.
3. Complejidad: está estrechamente ligada con la velocidad y el coste. Lo ideal sería encontrar un método lo más sencillo posible que diese una medida de calidad igual a la percibida. En general, resultados más próximos a los obtenidos por un observador se consiguen al incorporar características propias del SVH, lo que forzosamente implica que el método sea complejo.
4. Portabilidad: los resultados que proporciona el método no deben alterarse si se repiten las medidas en diferentes entornos o tiempos.
5. Precisión: referida a como representa el resultado del método de medida la percepción de calidad que tendría el observador.
6. Robustez: se pretende obtener resultados válidos sobre un amplio margen de variación de los parámetros asociados a la medida (tipo de

imagen, tipo de distorsión, condiciones de visibilidad, etc.), es decir, se buscan métodos robustos.

7. Forma del resultado: pueden ser valores numéricos (índices de calidad) o mapas de visibilidad del error. Según la aplicación a la que esté destinado un método de medida será conveniente una forma u otra y en general, lo ideal es que el método pueda proporcionar ambas salidas.

2. Objetivos del proyecto

El objetivo del proyecto es:

1. Investigar diferentes métricas de calidad de video streaming, tanto objetivas como subjetivas.
2. Implementar un modelo en Matlab para evaluar la calidad de video streaming sobre una red de fibra óptica.
3. Investicar la correlación entre las diferentes tecnicas.
4. Evaluar la bondad de los métodos seleccionados.

3. Métricas de Calidad Subjetivas

Los modelos subjetivos son aquellos que obtienen un valor como resultado de la evaluación de unos observadores a los que se les presentan los videos. Estos test no proporcionan información espacial o temporal. Los observadoes dan un valor de calidad global sin tener en cuenta donde y cuando aparecen los errores.

Los test subjetivos implican un elevado coste en tiempo y un numero elevado de personas. En primer lugar, la sala donde son llevados a cabo los test tiene que estar equipada de acuerdo con unas recomendaciones, y debemos realizar el test tantas veces como sea posible para obtener resultados fiables. El número de observadores y sus características (edad, profesión, experiencia, etc.) puede tener una influencia en los resultados. Para evitar esta dependencia debemos elegir un amplio y heterogéneo número de observadores.

Una recomendación ITU presenta un problema importante, llamado “efecto de contexto”. Es debido al orden e intensidad con que las distorsiones se presentan en el video. Por ejemplo, después de varios videos con pequeñas distorsiones si hay un video con una distorsión bastante notable el observador lo va a puntuar con peor puntuación que de lo que lo haría bajo otras circunstancias.

Aunque estos test conllevan algunos inconvenientes y tenemos que tener en cuenta que habrá un margen de error en los resultados, los test subjetivos son los más fiables para evaluar la calidad de video ya que las personas son los receptores finales en la mayoría de las aplicaciones streaming.

4. Métricas de Calidad Objetivas

Este tipo de métricas presenta numerosas ventajas: simplicidad, velocidad y lo más importante, el valor obtenido no depende de características especificas a la hora de tomar las medidas – características de un monitor, distancia al observador, etc. ni tampoco de observadores. Los resultados no se ven alterados si repetimos la misma medida otra vez.

Algunas de las métricas son simplemente formulas matemáticas y otras incluyen parámetros del Sistema Visual Humano que hacen la métrica más compleja pero a la vez más precisa.

A continuación explicare brevemente las técnicas objetivas simuladas en que he simulado en Matlab para evaluar la calidad de video streaming:

- SSIM: Structural Similarity Index

SSIM es un método de medida de la similitud entre dos imágenes basado en información estructural. Se basa en la suposición de que el sistema visual humano esta adaptado a extraer información estructural de las escenas de un video. Por ello esta basado en la degradación de esta información estructural, asumiendo que un error en el frame no es equiparable a perdida de calidad ya que algunas distorsiones pueden ser claramente visibles pero no molestas o viceversa.

Esta métrica se puede calcular sobre varias ventanas de un frame o una imagen. La medida entre dos ventanas 'x' e 'y' de tamaño NxN se obtiene según la fórmula (7).

Entre sus ventajas podemos destacar simplicidad, portabilidad, velocidad y bajo coste computacional. Además, proporciona no solo un valor global sino un mapa de calidad espacial.

El índice SSIM es un valor decimal entre 0 y 1. Un valor de indica que no existe ninguna correlación con el video original y un 1 significa que no ha habido degradación y se trata del mismo video.

- UQI: Universal Quality Index

Esta medida matemática, al igual que SSIM, no solo toma en consideración los errores o la distorsión en un video sino también como de visibles son estos en un video en particular.

Se le llama índice de calidad universal debido a que la medida no depende del tipo de imágenes que se están calificando o de las condiciones de los observadores. Más importante, puede aplicarse en diferentes tipos de aplicaciones con diferentes tipos de distorsiones.

El índice UQI se define según la formula (8). El rango dinámico de UQI es entre -1 y +1. Este indice de calidad modela cualquier distorisión como combinación de tres factores diferentes: perdida de correlación, luminancia y contraste.

- SC: Structural Content

Valores elevados de SC implican que la imágen o el video tiene calidad pobre. SC es una metrica definida según la formula (3)

- MD: Maximum Difference

Valores elevados de MD implican pobre calidad en el video o imagen. MD se define como en la formula (4)

- LMSE: Laplacian Mean Square Error

Esta es una medida basada en la importancia de los bordes. Un valor elevado de LMSE significa mala calidad del video. Se define según (5)

- NAE: Normalized Absolute Error

Valores elevados de NAE indicant pobre calidad de imagen o video. NAE se define según (6)

- MSE: Mean Square Error

Una de las medidas más simples de calidad de imagen es MSE. Valores elevados de MSE significan pobre calidad de imagen. MSE se calcula como en la formula (1). Cuando dos imágenes son idénticas MSE será igual a cero.

- PSNR: Peak Signal to Noise Ratio

La medida más común de todas las métricas objetivas es PSNR calculada según (2). PSNR mide la fidelidad con que un frame o video se parece al original.

Comparando con otras medidas objetivas, PSNR es fácil de obtener y sencilla. El mayor problema es que solo tiene en cuenta la luminancia, y no tiene en cuenta los componentes de crominancia, que es importante para la percepción humana.

Debido a que en ningún caso tenemos en cuenta plenamente la sensibilidad del sistema visual humano todas las distorsiones son tomadas con la misma relevancia, sin importar de que tipo son, su localización o escena en la que suceden. Es por esto por lo que necesitamos probar correlación entre los resultados de medidas objetivas y subjetivas.

5. Implementación

He recibido 38 muestras de video de mi compañero de proyecto que corresponden a dos diferentes videos, uno con imágenes más estáticas y otro con mayor tasa de movimiento. Estos videos fueron comprimidos con MPEG-2 y después se realizó una simulación en Matlab de transmisión a través una red fibra óptica. Los videos son el resultado de una transmisión sobre una red con mucho trafico y con diferentes parámetros de red – tasa de perdida de paquetes, delay y jitter.

He implementado en Matlab las métricas objetivas arriba mencionadas según las formulas y he obtenido los resultados componente a componente del video y frame a frame. A continuación he calculado la media para cada muestra de video que es el valor global final.

Una observación importante es que no todos los parámetros que afectan a la calidad de video se pueden considerar por igual. El efecto de la tasa de frame, por ejemplo, porque en las métricas utilizadas necesitamos el video original y comparamos frame a frame. Esto significa que ambos videos, el original y el distorsionado, deben tener el mismo número de frames. Si no los resultados no serán reales sino degradados.

A continuación he usado el programa externo ‘bvqm_pc_v12’ para obtener los valores de calidad VQM para cada muestra de video y he guardado todos los resultados en archivos *.mat.

La bondad de una métrica objetiva se evalúa hallando la correlación entre resultados objetivos y los test subjetivos. Los últimos suelen ser los llamados MOS (Mean Opinion Scores).

Para analizar la correlación entre los diferentes algoritmos de calidad y entre cada uno de ellos y resultados cercanos a resultados subjetivos he usado VQM en lugar de MOS (Mean Opinion Score), debido a que es un algoritmo automático que nos proporciona valores próximos a MOS- tienen un valor de correlación de Pearson de 0.9- De esta forma no es necesario llevar acabo test subjetivos y no evitamos todos los inconvenientes de MOS.

A continuación he implementado las medidas de correlación mas usadas para analizar las métricas de calidad de video, estas son: coeficiente de correlación lineal, correlación de Spearman, coeficiente Kappa y kurtosis.

6. Coeficiente de Correlación Lineal

El valor denominado coeficiente de correlación lineal mide la fuerza y dirección de la relación entre dos variables. El valor de este coeficiente puede ser $-1 \leq r \leq +1$. Los símbolos + y – indican que la correlación es bien positiva o bien negativa. Valores positivos entre una variable ‘x’ y otra variable ‘z’ indica que para valores crecientes de ‘x’ tendremos valores crecientes de ‘z’ y viceversa. Si ‘x’ y ‘z’ tienen una correlación fuerte el valor es cercano a 1. Un valor exacto de 1 indica que ambas métricas coinciden a la perfección. Una correlación mayor a 0.8 se describe generalmente como fuerte mientras que menor de 0.5 se describe como débil.

Como podemos observar en la tabla 5 las métricas más relacionadas con VQM son NAE (normalized absolute error), PSNR(peak signal to noise rate) y SSIM(structural similarity index). UQI(universal quality index), MSE(mean square error) y SC(structural contex) tienen también una fuerte correlación mientras que MD(maximum difference) and LMSE (Laplacian Mean Square Error) muestran correlación débil. También podemos observar que MD tiene una correlación débil con todas las otras métricas.

7. Coeficiente Spearman

El coeficiente de Spearman es una medida de la relación lineal entre dos variables. Se diferencia del coeficiente de Pearson en que los cálculos se realizan después de que los resultados de las medidas se convierten en rankings. De esta forma he convertido todos los resultados de la tabla 4 en valores de 1 a 5, dando un 1 a los mejores resultados y un 5 a los peores. Aquellas métricas que tienen correlación inversa con VQM tienen ranking inversos -SSIM, UQI, PSNR, MD.

Como en el coeficiente de correlación lineal el valor de Spearman puede ser entre -1 y +1. Cero implica ninguna correlación mientras que ± 1 significa que tenemos correlación perfecta.

Con Spearman se han obtenido resultados similares que con Pearson como se puede ver en la tabla 6. Los mejores resultados son obtenidos con NAE, PSNR y SSIM. UQI y SC tienen también correlación fuerte con VQM y las relaciones más débiles son VQM-MSE y VQM-MD.

8. Coeficiente Kappa

Para obtener el coeficiente Kappa hay que trabajar también con rankings. Kappa proporciona una medida del acuerdo entre dos evaluaciones, dos métricas diferentes en este caso, de si coinciden en sus clasificaciones de la calidad de videos en puntuaciones excluyentes, en este caso 5: excelente calidad (1), buena, suficiente, pobre y mala (5). Esta medida toma en consideración el porcentaje de acuerdo entre observadores y el porcentaje de acuerdo al azar.

Los mejores resultados han sido obtenidos de nuevo con NAE (perfect agreement), SC (moderate agreement), SSIM (fair), PSNR (fair). Los peores resultados han sido obtenidos con MD, LMSE y MSE (slight agreement).

9. Kurtosis

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean.

Se ha usado la función de Matlab `kurtosis()`, y a continuación comparo la kurtosis de los resultados de cada métrica para todos los videos con la kurtosis de VQM.

Si las métricas tienen un alto valor de kurtosis podemos interpretar que es un método bastante sensible a variaciones en parámetros de la red, ya que cada video es el resultado de una simulación con parámetros diferentes.

SSIM y UQI tienen resultados similares como en las gráficas 2-3. LMSE obtiene similares resultados para los dos tipos de videos. También UQI y SSIM. SC y MSE no reconocen tantas perdidas como las otras dos métricas en las muestras de video estático pero obtienen resultados bastante buenos con las muestras de video con frames con mayor movimiento. Por otra parte PSNR, MD y VQM obtienen buenos resultados.

10. Conclusiones

Las conclusiones de este estudio sobre el efecto del jitter y deley en la calidad de video son:

- Los efectos de jitter en la degradación de la calidad del video percibido son muy similares a los efectos de las perdidas. Es mas, una degradación puede ser severa con niveles bajos de perdidas y/o jitter, mientras que valores más elevados de los mismos parámetros no degradan la calidad de forma proporcional.
- Como era de esperar, aquellas secuencias más estáticas-con menor variación temporal- son más robustas frente a la presencia de jitter que aquellas con mayores niveles de variación temporal.

Las conclusiones de este estudio sobre la bondad de los algoritmos de estimación de calidad:

- La robustez del tipo de métrica es un parámetro importante; necesitamos tener resultados validos en un amplio rango de parámetros relacionados con la métrica- el tipo de video y el tipo de distorsión. SSIM y UQI parecen los métodos más robustos, este puede ser debido a que no solo toman en consideración errores de distorsión sino también como de visibles son en una imagen en particular.
- La mayor fiabilidad de medidas de calidad objetivas para videos comprimidos con MPEG-2 han sido obtenidos con NAE y PSNR. Estas son las métricas simples más adecuadas y las que recomiendo para calidad de video streaming en redes de fibra óptica. Como se puede comprobar, la relación entre medidas objetivas y medidas VQM en el video comprimido mostradas en las gráficas 8-4 respectivamente parecen más cercanas a una correlación lineal que el resto.
- Las métricas con menor correlación con VQM fueron LMSE, MSE y MD. Como se puede observar, las gráficas de LMSE, MSE y MDE tienen puntos más esparcidos.

8. REFERENCES

[1] ITU-T Rec. J.247 (08/08) – “Objective perceptual multimedia video quality measurement in the presence of a full reference“

[2] Alan Clark, Ph.D., “Clarifying Video Quality Metrics“-http://news.tmcnet.com/ - April 11, 2006

[3] Ulrich Engelke and Hans-J¨urgen Zepernick- “Perceptual-based Quality Metrics for Image andVideo Services: A Survey”

[4] Eckert, M.P.; Bradley, A.P. “Perceptual quality metrics applied to still image compression”, Signal Proc.,

[5] Ratchakit Sakuldee, and Somkait Udomhunsakul, “Objective Performance of Compressed Image Quality Assessments” - Procedings of World Academy of Science, Engineering and Technology Volume 26 December 2007 ISSN 2070-3740

[6] ITU-T Recommendation P.910 - “Subjective video quality assessment methods for multimedia applications”

[7] Stephen Wolf, Margaret H. Pinson, Arthur A. Webster,Gregory W. Cermak, E. Paterson Tweedy - “Objective and Subjective Measures of MPEG Video Quality” Institute for Telecommunication Sciences Boulder, CO; GTE Laboratories Incorporated Waltham, MA

[8] Ana Rosa Gallego Aguilar – “Modelos visuales en el análisis de
La calidad de imagen” - February 2006, Universidad Politecnica de Madrid

[9] G.J. Sullivan and T. Wiegand: Video Compression - from concepts to the H.264/AVC standard, *Proceedings of the IEEE*, Vol. 93, No. 1, Jan. 2005, pp. 18-31

[10] http://www.iptvtroubleshooter.com/video_quality.html - February 09.

[11] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity” - IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 13, NO. 4, APRIL 2004

[12] Erik Valdemar Cuevas Jimenez, Daniel Zaldivar Navarro, “Visión por Computador utilizando MatLAB Y el Toolbox de Procesamiento Digital de Imágenes”

[13] Karen Egiazarian, Jaakko Astola, “Two New Full-Reference Quality Metrics Based on HVS” Tampere University of Technology, Tampere, Finland

[14] A. Bovik, “Handbook of Image and Video Processing,” Academic Press, 2000.

[15] Muhammad Saqib Ilyas, “Perceptual Video Quality Asessment” - Lahore University of Management Sciences

[16] Patrizio Campisi, Patrick Le Callet, and Enrico Marini, “Stereoscopic Images Quality Assesment”- Dipartimento di Elettronica Applicata, Università degli Studi di Roma

[17] VQEG, Video Quality Expert Group, www.its.bldrdoc.gov/vqeg

[18] ITU-R BT.500-11- “Methodology for the Subjective Assessment of the Quality of Television Pictures”.

[19] Cadik, M.; Slavik, P, “Evaluation of two Principal Approaches to Objective Image Quality Assessment”- Eight International Conference on Information Visualisation (IV’04)

[20] Zhou Wang, and Alan C. Bovik – “A Universal Image Quality Index” - IEEE SIGNAL PROCESSING LETTERS, VOL. 9, NO. 3, MARCH 2002 81