QUALITY ASSESSMENT OF CHROMATIC VARIATIONS: A STUDY OF FULL-REFERENCE AND NO-REFERENCE METRICS

Marco V. Bernardo†*, *, *António M. G. Pinheiro*†*, *Paulo T. Fiadeiro*†*, *Manuela Pereira**

†Remote Sensing Unit
*Instituto de Telecomunicações
Universidade da Beira Interior
Rua Marquês d'Avila e Bolama - 6201-001 Covilhã, Portugal
{mbernardo, pinheiro, fiadeiro, mpereira}@ubi.pt

ABSTRACT

This work describes a comparative study on the ability of Full-Reference *versus* No-Reference quality metrics to measure the Quality of Experience created by images that suffer chromatic variations. Considering this, some well known Full-Reference (PSNR, UQI, MSSIM) and No-Reference (GM, FTM, RTBM) will be compared with the MOS results. Although the quality metrics considered are usually applied to the luminance component, in this study they are applied to the Y, C_b, C_r components separately. The result of the three components average metrics was also considered, because only the image chromatic components have been changed resulting in similar values of luminance. The correlation estimates show that the Full-Reference Metrics namely the MSSIM and the UQI provide a good representation of the subjective results. Moreover, the studied No-Reference metrics also provide an acceptable representation, although their reliability is less effective.

***Index Terms*—** Quality of Experience, Mean Opinion Score, Quality Metrics, Image Quality

1. INTRODUCTION

Nowadays, analyzing the performance of a multimedia system requires an analysis of the Quality of Experience (QoE). As QoE involves the subjective factors of the user it results in larger reliability performance evaluation of the system performance [1]. Although subjective assessment of audio and visual quality is very expensive and time consuming, it is considered to be the most accurate method to reflect the human perception [2]. However, considering the new requirements of industry, the huge variety of systems, and also the large numbers of content providers and users, becomes impossible to rely only in the subjective assessment. A number of objective methods for measuring the perceived video quality have been proposed for objective video quality assessment [3].

The current work relates the influence of chromatic variation in the QoE with common Full-Reference (FR) and No-Reference (NR) objective metrics, complementing the analysis of the influence of the chrominance information errors [4]. This study will use a set of predefined FR and NR metrics, selected because of their performance after initial testing using a large number of state of the art metrics [5]. In particular, the FR metrics Peak Signal-to-Noise Ratio (PSNR), Universal Image Quality Index (UQI), Mean Structural SIMilarity index (MSSIM) and NR metrics Gradient Metric (GM), Frequency Threshold Metric (FTM) and Riemannian Tensor Based Metric (RTBM), are compared with the Mean Opinion Score (MOS). Their performance is evaluated using three prediction attributes: accuracy, monotonicity and consistency.

Chrominance is one of the four International Telecommunication Union (ITU) indicators for spatial distortion analysis of the perceptual evaluation of visual quality. The color gamut of natural scenes is constrained to the more central region of the chromaticity diagram. To relate color-vision research in the color gamut with the image quality assessment is an important issue. In this context, any chromatic distortion may affect the image appearance, especially when natural images, acquired by hyperspectral imaging systems, are used as the stimuli. The availability of hyperspectral images of natural scenes with high chromatic resolution has made possible the extension of gamut-mapping analysis to real natural stimuli.

The hyperspectral image data used in previous study [4] was obtained from two databases: three natural scenes (rural and urban scenes) from University of Manchester hyperspectral image database [6] and two urban environments from a recently created hyperspectral image database at Universidade da Beira Interior and Universidade de Coimbra. This set of images with induced chromatic variations and with available MOS values is used for testing.

2. PREVIOUS WORK

In previous work [4], the sensibility to chromatic variations was tested and quantified using subjective testing. In [7] was tested the use of FR metrics. In [4] the MOS was computed,

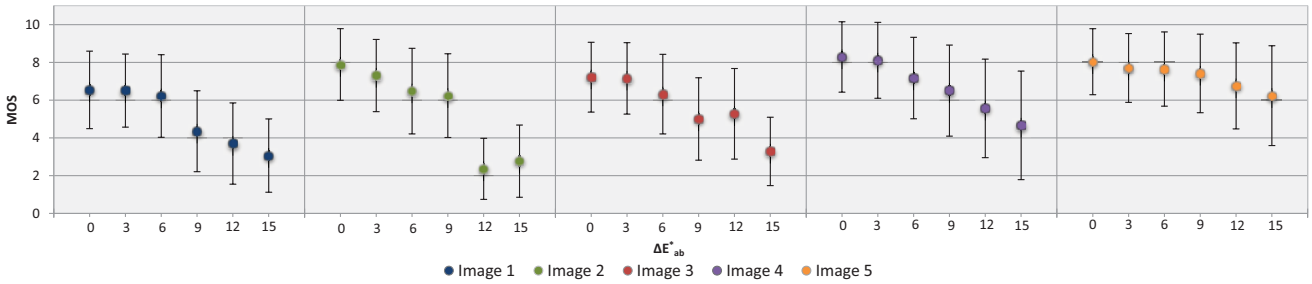

Fig. 1. MOS as function of ΔE^*_{ab} for all test images.

allowing to test and quantify the sensibility to chromatic variations. To obtain real images of colored natural scenes with different chromatic variations it was necessary to manipulate the spectral reflectance data from a database of hyperspectral images. These images were then converted into spectral radiances using a $D65$ illuminant to obtain the corresponding representation in the CIE 1976 ($L^*a^*b^*$) color space and a true color image representation system was defined. This color space, also known as CIELAB, is device independent, partially uniform and based on the Human Visual System (HVS) [8–11].

To avoid spatial artifacts, the colors of the images were subdivided into clusters applying the K-Means algorithm [12]. Then any color pixel i of an image, represented by $(L^*_i a^*_i b^*_i)$, that belong to a specific cluster were chromatically transformed into a new color pixels $(L^*_{ei} a^*_{ei} b^*_{ei})$ by adding a ΔE^*_{ab} error (equation (1)) with a predefined magnitude and random direction.

$$\Delta E^*_{ab} = \sqrt{(L^*_i - L^*_{ei})^2 + (a^*_i - a^*_{ei})^2 + (b^*_i - b^*_{ei})^2} \quad (1)$$

This procedure was applied to all color clusters keeping the chromatic error, but applying different random directions. Hence, was guaranteed that groups of similar colors were changed in the same direction. Also, the evaluation provided will depend exclusively on the chromatic variations, which was the aim of the study, because color artifact have been reduced to unperceived levels. The magnitude of ΔE^*_{ab} ranged from 3 to 15 units in steps of 3 units. A set of 5 different images were generated for each predefined magnitude error to cover a larger number of directions in the CIELAB color space.

The subjective quality assessment experiment was conducted at a compliant laboratory that follows the recommendations for subjective evaluation of visual data issued by ITU-R [13] and was chosen the Single Stimulus Continuous Quality Evaluation (SSCQE) standard test methodology.

Figure 1 presents the MOS as a function of the chromatic error ΔE^*_{ab} for the test images grouped by the same chromatic error [4]. As expected, the MOS value decrease with the increase of the chromatic error ΔE^*_{ab} for all images.

3. OBJECTIVE QUALITY ASSESSMENT

Objective Image Quality Metrics (IQMs) can be classified as FR, Reduced-Reference (RR) and NR. The FR image quality assessment requires the reference image prior to any distortion. For the RR, the reference image is not available, but it is represented by a set of extracted features representative of the image quality. Finally, the perceived quality is computed in the absence of the original image for the NR approaches.

This work is focused on a comparative study between FR and NR metrics. These are common metrics in the related literature and represent an approach to the perceptual quality. They have been chosen for this reason, and also because they provided the best performance after some preliminary testing. Moreover, they will be compared with the MOS obtained in the previous study [4]. All these metrics are typically applied to the luminance channel. In this study the original color images were converted into the YC_bC_r color space and these metrics were applied to the components Y, C_b, C_r separately. The average of the three components Y, C_b, C_r metrics were also calculated for the objective quality assessment. The FR metrics are generally more reliable as they provide a comparison to the reference image, while the NR metrics might be very usefull to estimate the image quality in the absence of original image.

3.1. Full-Reference Metrics

The studied FR assessments can be divided in two different categories. The first includes difference based measures like Mean Squared Error (MSE) or the PSNR. The second is based on the HVS and includes structure similarity measures like the UQI [14] or the MSSIM [15].

The image and video processing community has been using the MSE and PSNR as fidelity metrics for a long time. These two metrics are quite popular since their computation is simple and fast.

The MSE represents the power of the difference between original and distorted images. PSNR is just a logarithmic representation of the MSE. It is usually expressed in terms of the logarithmic decibel where 255 is the maximum possible amplitude for an 8-bit image. High values of the PSNR, define

an improved reproduction quality. Although PSNR does not provide a perceptual visual quality measure, as it is based on a purely pixel by pixel difference measure [16], it has been widely used.

The UQI metric was proposed by Wang and Bovik [14]. Instead of using traditional error summation methods, this metric was designed to model any image distortion as a product of three factors: loss of correlation, luminance distortion and contrast distortion. Due to the combination of these three factors, the metric is independent of visualization conditions or individual observers. Additionally, it is also easier to calculate the metric value bacause of the low complexity. The authors indicate that UQI results are significantly better than MSE and PSNR for different types of distortion. Measures can have values in the range $[-1, 1]$, where 1 represents the comparison of two identical images.

The Structural SIMilarity (SSIM) metric is an improved version of the UQI. It is a perceptual metric based on the content features extraction and abstraction. This quality metric considers that the HVS uses the structural information from a scene [15]. The structure of the objects in the scene, can be represented by its attributes, which are independent of both contrast and average luminance. Hence, the changes in the structural information from the reference and distorted images, can be perceived as a measure of the image distortion. MSSIM value is the mean value of SSIM map over the whole image and can have values in the range $[0, 1]$, where 1 represents the comparison of two identical images.

3.2. No-Reference Metrics

The studied NR assessments were based on the analysis of a few well known sharpness measures. These metrics were selected after preliminary studies with a larger set of NR metrics. Gradient Metric (GM) [17], Frequency Threshold Metric (FTM) [5] and Riemannian Tensor Based Metric (RTBM) [18] were tested and their results compared with the available MOS results.

The image gradient is quite common in different applications because of its computation simplicity and effectivity. The GM is based on the study by Batten [17] and it results from the mean image gradient. It is based in the well known concept that the gradient has a stronger response in areas of significant gray-level transitions because it highlights the regions limits.

Firestone *et al* [19] describes a method using spectral analysis for optical microscopy, also applicable to electron microscopy. This spectral analysis was applied to the evaluation of the images focus. When the image becomes focus, sharpen edges and fine details become more visible since it corresponds to high spatial frequencies. Some differences to the method were introduced by Murthy *et al* [5] and the FTM corresponds to the addition of all the magnitudes for the frequency components different from zero, below a certain threshold.

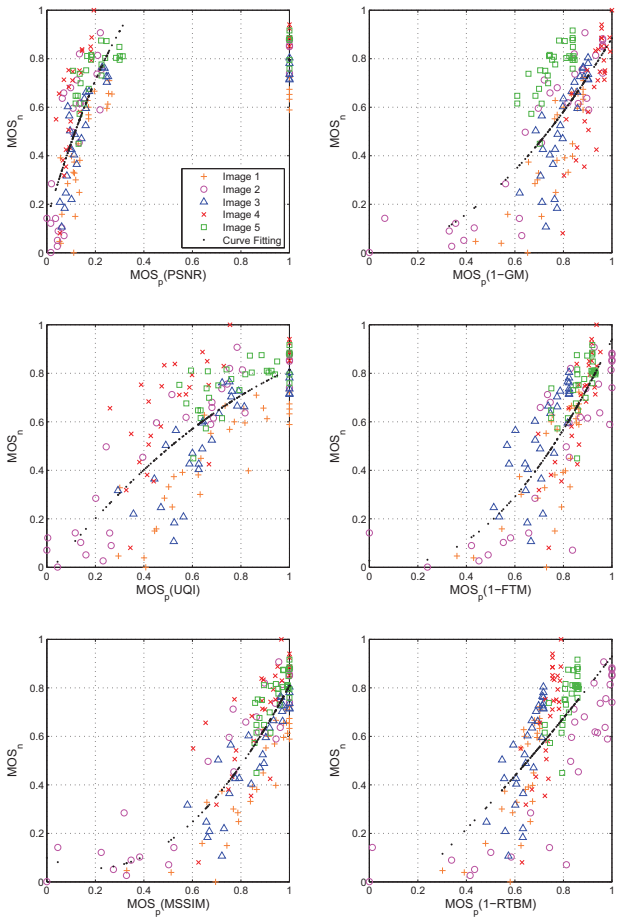

Fig. 2. Fitting analysis for MOS vs IQMs using the three components Y, C_b, C_r average.

Finally, the RTBM is another NR objective metric based on the sharpness. It uses the Riemannian tensor by mapping the image into a non-Euclidean space and measuring the curve variation [18]. It was shown that this metric predicts the perceived sharpness even in the presence of noise. The result of this metric is the inner product between the tangent vector to the Riemannian manifold and the manifold itself.

All of these six IQMs are compared with the MOS available for the studied images and the ability to provide a representation of the QoE was assessed. The statistical relation between the MOS, FR and NR metrics will be presented in the next section.

3.3. Evaluation of Objective Models

The original subjective results (MOS) were normalized into MOS_n on the range $[0, 1]$, shown in equation 2, where i is the i^{th} subjective test.

$$\text{MOS}_n(i) = \frac{\text{MOS}(i) - \text{MOS}_{min}}{\text{MOS}_{max} - \text{MOS}_{min}} \quad (2)$$

Then, the non-linear regression suggested in [20], was fitted to the IQMs and the mapped MOS_n values, and restricted to be monotonic over its range. The equation (3) fitted to the data $[\text{MOS}_p, \text{MOS}_n]$ was used in the regression, where MR is the metric result, and $b1$, $b2$, and $b3$ denote the regression parameters.

$$\text{MOS}_p = \frac{b1}{1 + e^{-b2 \times (MR - b3)}} \quad (3)$$

Figure 2 illustrates the analysis of fitting the MOS *vs* IQMs for all images, using the average of the three components Y, C_b, C_r metrics. The obtained expressions for fitting IQMs for all images are shown in Table 1.

Table 1. Obtained expressions for IQMs fitting.

IQM	Fitting curve expressions
PSNR	$-2.752x^2 + 3.408x + 0.139$
UQI	$-0.392x^2 + 1.237x - 0.031$
MSSIM	$1.159x^2 - 0.449x + 0.099$
GM	$0.759x^2 - 1.663x + 0.884$
FTM	$1.180x^2 - 2.091x + 0.939$
RTBM	$0.219x^2 - 1.319x + 0.930$

To evaluate the metrics performance four measures were chosen [21]. 1) The Pearson linear correlation coefficients between MOS_n and MOS_p, that measures the model prediction accuracy. 2) The Spearman rank order correlation coefficient between MOS_n and MOS_p, that evaluates the model prediction monotonicity. 3) The Outlier Ratio as a measure of the model consistency prediction. 4) The root mean square error (RMSE).

Table 2. Evaluation results of IQMs for all the test images using the average of the three components Y, C_b, C_r metrics.

IQM	Pearson	Spearman	Outlier ratio	RMSE
PSNR	0.500	0.759	0.013	0.395
UQI	0.754	0.721	0.020	0.186
MSSIM	0.811	0.791	0.027	0.305
GM	0.705	0.646	0.013	0.269
FTM	0.774	0.800	0.013	0.265
RTBM	0.723	0.746	0.013	0.221

Table 2 presents the performance measures for the different IQMs studied in this work. These measures were computed for all metrics PSNR, UQI, MSSIM, GM, FTM and RTBM considering the average of the three components Y, C_b, C_r metrics.

For the FR metrics the MSSIM results in the highest performance measures, closely followed by the UQI metric. The MSSIM has a slightly better Pearson correlation for all images. The outlier ratio is very small for all IQMs. Finally, the UQI results better for RMSE. For the NR metrics the FTM results in the highest performance measures, closely followed by the RTBM metric. Furthermore the NR metric FTM has a similar performance to the best FR metrics.

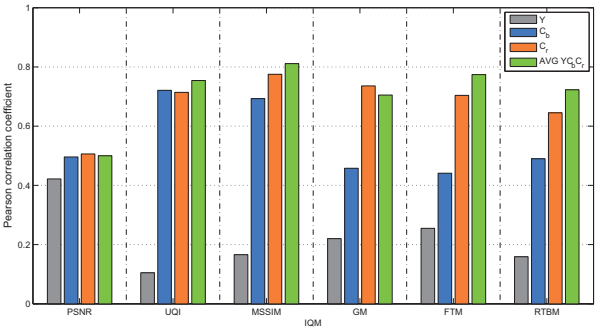

Fig. 3. Pearson analysis.

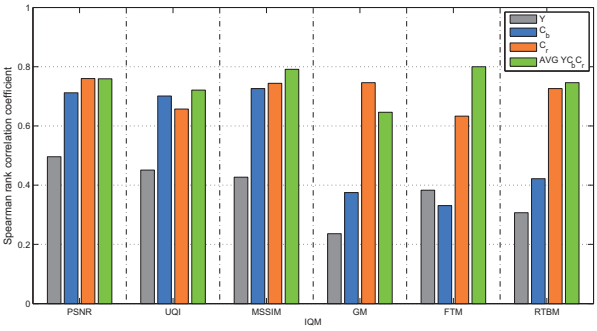

Fig. 4. Spearman analysis.

The values of the Pearson and Spearman correlation are also represented by the bar plots in figures 3 and 4, respectively. In these plots can be observed that the results computed using the average of the three components Y, C_b, C_r metrics result in a better evaluation than the metrics computed over the individual components for the FR metrics.

4. CONCLUSION

In this work the ability of six IQMs to provide a good representation of the subjectives tests represented by the MOS of color images with applied of chromatic errors was studied. The FR metrics MSSIM and the UQI and the NR metrics FTM and the RTBM, provide a representation with acceptable reliability. Moreover, the FTM performance is similar to the best FR metrics. Furthermore the metrics result in higher reliability using the average of the three components Y, C_b, C_r metrics.

REFERENCES

[1] Peter Reichl, Joachim Fabini, Marco Happenhofer, and Christoph Egger, "From QoS to QoX: A Charging Perspective.," in *Proc. 18th ITC Specialist Seminar on Quality of Experience, Blekinge Institute of Technology, Karlskrona*, May 2008.

[2] ITU, "ITU-T Recommendation P.911: Subjective audiovisual quality assessment methods for multimedia application," Tech. Rep., ITU - Telecom. Standard. Sector, December 1998.

[3] ITU, "ITU-T Recommendation J.144: Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," Tech. Rep., ITU - Telecom. Standard. Sector, March 2001.

[4] Marco V. Bernardo, António M. G. Pinheiro, Manuela Pereira, and Paulo Torrão Fiadeiro, "A study on the user perception to color variations," in *Proceedings of the 20th ACM international conference on Multimedia*, Nara, Japan, 2012, pp. 1009–1012.

[5] A.V. Murthy and L.J. Karam, "A MATLAB-based framework for image and video quality evaluation," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, June 2010, pp. 242–247.

[6] D. Foster, S. Nascimento, and K. Amano, "Information limits on neural identification of colored surfaces in natural scenes.," *Visual Neuroscience*, vol. 21, no. 3, pp. 331–336, 2004.

[7] Marco V. Bernardo, António M. G. Pinheiro, Manuela Pereira, and Paulo Torrão Fiadeiro, "Objective evaluation of chromatic quality assessment," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, July 2013, pp. 1–6.

[8] CIE, "Colorimetry official recommendation of the international commission on illumination," CIE publication 15.2, CIE Central Bureau, 1986.

[9] T. Smith and J Guild, "The C.I.E. colorimetric standards and their use," *Transactions of the Optical Society*, vol. 33, no. 3, pp. 73–134, 1931.

[10] R. Hunt, *The Reproduction of Colour*, John Wiley & Sons, New York, 3 edition, March 1975.

[11] G Wyszecki and W Stiles, *Color Science*, John Wiley & Sons, New York, 3 edition, March 1967.

[12] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall PTR, 2011.

[13] ITU, "ITU-R Recommendation BT 500-12: Methodology for the subjective assessment of the quality of television pictures," Tech. Rep., ITU - Radiocom. Sector, September 2009.

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

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

[16] Bernd Girod, "What's wrong with mean-squared error?," in *Digital images and human vision*, Andrew B. Watson, Ed., pp. 207–220. MIT Press, Cambridge, MA, USA, 1993.

[17] Christopher F. Batten, "Autofocusing and astigmatism correction in the scanning electron microscope," M.S. thesis, University of Cambridge, U.K., 2000.

[18] Rony Ferzli and Lina J. Karam, "A no-reference objective sharpness metric using riemannian tensor," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, January 2007.

[19] Lawrence Firestone, Kitty Cook, Kevin Culp, Neil Talsania, and Kendall Preston, "Comparison of autofocus methods for automated microscopy," *Cytometry*, vol. 12, no. 3, pp. 195–206, 1991.

[20] ITU, "ITU-T Tutorial - Objective perceptual assessment of video quality: full reference television," Tech. Rep., ITU - Telecom. Standard. Sector, 2004.

[21] VQEG, "Final report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment," Phase II VQEG, March 2000.