IMPROVING FULL-REFERENCE IMAGE QUALITY ASSESMENT USING MACHINE LEARNING

Martin D. Dimitrievski
Faculty of Electrical Engineering and
Information Technologies
Skopje, Macedonia

Zoran A. Ivanovski, *IEEE Member*
Faculty of Electrical Engineering and
Information Technologies
Skopje, Macedonia

ABSTRACT

In this work we comprehensively analyze three standard image quality measures, measure their performance on gray scale images with different qualities against opinions from human viewers, and propose a novel image quality measure. Furthermore we inspect the behavior of several types of image degradations and the ability of each metric to detect the quantity of perceived loss of image quality for each degradation type. It was found that these measures based on first and second order statistics, although computationally simple, do not correlate very well to mean opinion scores. By means of machine learning we find an optimal regression model that combines these statistical metrics into a novel full reference image quality metric which correlates better to human scores for all tested degradation types.

I. INTRODUCTION

Image quality metrics (IQM) are quantities used for the evaluation of imaging systems or of coding/processing techniques. In this study we consider several image quality metrics and study their statistical behavior when measuring various compression or sensor artifacts. A good objective quality measure should reflect the distortion on the image well due to blurring, noise, compression etc. It is expected that such measures could be instrumental in predicting the performance of vision-based algorithms such as feature extraction, image based measurements, detection or tracking. In the image coding and computer vision literature, the most frequently used metrics are differences between the original and coded images, [1-3] with varieties of the mean square error (MSE) or signal to noise ratio (SNR) as the most common ones. The reasons for their widespread popularity are their mathematical tractability and the fact that it is often straightforward to design systems that minimize the MSE. However they are not very well matched to perceived visual quality [4]. Concluding that natural image signals are highly structured authors of the Structural Similarity Index Metric (SSIM) in [5] are motivated to find a more direct way to compare the structures of the reference and the distorted signals. Their paradigm is a top-down approach, mimicking the hypothesized functionality of the overall human visual system (HVS) as a particular implementation of the philosophy of structural similarity, from an image formation point of view. This metric too has its limitations when compared to subjective scores due to its simple model. Another standard image quality assessment (IQA) metric is the Universal Quality Index (UQI) proposed in [6]. This metric is mathematically defined and although it is not based on any HSV model it shows significantly better on various distortions than standard peak signal to noise ratio (PSNR).

Our observation as presented in the following sections shows that each of the three metrics has its strengths when measuring degradations quantities of some types but fail in other cases. We explore the idea of fusing different features using ε-SV regression and test the accuracy of the final vote against all degradation types from the LIVE2 database [7]. For regression we use the LIBSVM implementation [8].

The remaining sections are organized as follows: In section II we discuss five different distortion types which happen frequently in digital images, section III gives a detailed analysis of each of the three statistical metrics discussed, using machine learning we present a novel metric approach and its performance in section IV and finally we bring a conclusion in section V.

II. IMAGE DEGRADATION TYPES

When speaking of image degradation we concentrate on distortions visible on the digital images to the end user. Due to the nature of efficient storage and transfer of digital media over the internet images are subject to loss of visual quality of different types. Storage of digital media is essentially performed by a lossy compression algorithm of the bit stream followed by statistical coding. Both of these processes introduce perceivable distortions when the stored image is decompressed and presented to the end user. Lossy compression algorithms although very efficient in reducing the redundancy of information in the bitsreams result in general loss of high frequency details and other artifacts such as blocking and ringing. Coder errors also contribute to the loss of perceived visual quality by reducing the fidelity of the color gamut.

Other class of degradation which occurs frequently and is very annoying to the viewer is the image noise. The very process of image acquisition relies on complex electronic circuitry for conversion of light to digital information. Due to thermal dissipation inside the imager light information is intermixed with ambient heat which results in reduction of the SNR. In most literature noise in digital images is modeled as an additive white Gaussian process to the image signal. This approximation results in a simple model which very closely explains the natural effect of image noise and is often used as a starting point for denoising algorithms.

Network communications are an inevitable system responsible to the quality of digital media. Inspired by the broad usage of wireless networks we also include a type of degradation which occurs when using an unreliable wireless channel for image transfer. Channel noise or fading may either be due to multipath propagation, referred to as multipath induced fading, or due to shadowing from obstacles affecting the wave propagation, sometimes referred to as

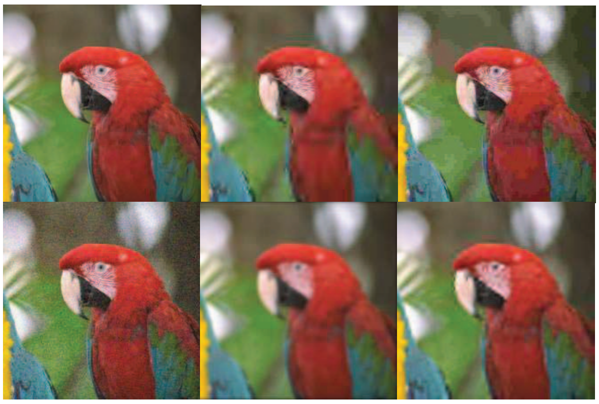

Figure 1. Parrots image distorted with 5 distortions. From left to right and from top to bottom: original image, fast fading, jpeg, noise, blur, jpeg2000

shadow fading. Fast fading occurs when the coherence time of the channel is small relative to the delay constraint of the channel. In this regime, the amplitude and phase change imposed by the channel varies considerably over the period of use. This results in general loss of data packets and thus parts of the transferred image. Fast fading channel is modeled with introducing bit errors to a JPEG2000 bit stream which results in loss of bands of spatial frequencies in the images. Figure 1 visualizes all abovementioned degradations on the same image.

In this paper we use the second release of the LIVE database [7] which contains 29 reference images degraded with various degradation processes. Each image is degraded with one of the following five degradation types:

1. JPEG compression
2. JPEG2000 compression
3. Additive Gaussian noise
4. Gaussian blur
5. Fast-fading channel noise

Individual types of degradations are applied several times with different strengths to every reference image which results in a dataset of 779 distorted images. For each image a mean opinion score (MOS) is computed out of the scores given as an expert knowledge or "ground truth" from interviewing many different users. MOS range from 0 to 100 quantifying the level of degradation or loss of perceived visual quality of the image where 0 means a low quality image and 100 is an image with no visible degradation.

III. USED METRICS

Before we start the analysis of the used metrics first we define a performance measure with which we measure how well the metric correlates to the human observer. Correlation is measured using the Spearman Rank-Order Correlation Coefficient (SROCC) and the Pearson product-Moment Correlation Coefficient (PMCC). The former estimates how well the relationship between the two variables can be described using a monotonic function and the later measures the linearity of the two variables.

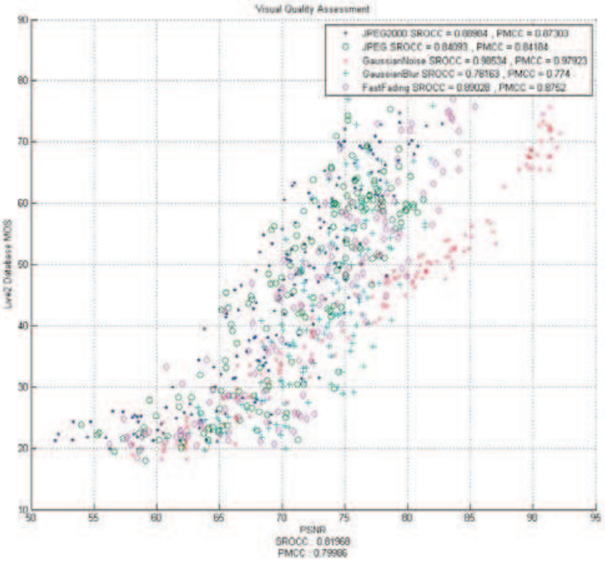

Figure 2. PSNR versus MOS for each degradation type

The most wide used metric which has been an industry standard for measuring the quality of compression codecs is a term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale. The signal in this case is the original data, and the noise is the error introduced by compression. When comparing compression codecs it is used as an approximation to human perception of reconstruction quality, therefore in some cases one reconstruction may appear to be closer to the original than another, even though it has a lower PSNR (a higher PSNR would normally indicate that the reconstruction is of higher quality). PSNR is defined through the MSE:

$$MSE(I,K) = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \qquad (1)$$

where I and J are the original and distorted image and m and n are the number of rows and columns.

$$PSNR(I,K) = 10\log_{10}\left(\frac{MAX_I^2}{MSE(I,K)}\right) \qquad (2)$$

where *MAX* is the maximum possible value of the image. For images represented with 8 bits per sample this value is 255. Typical values for PSNR in our dataset are ranging from 50 to 90dB where higher usually is better. In Figure 2 we present the PSNR of all 779 images from the LIVE database plotted against the mean opinion score for each type of degradation. SROCC and PMCC values for the entire dataset and for each degradation class are computed separately and are given in the legend of the figure and in Table 1.

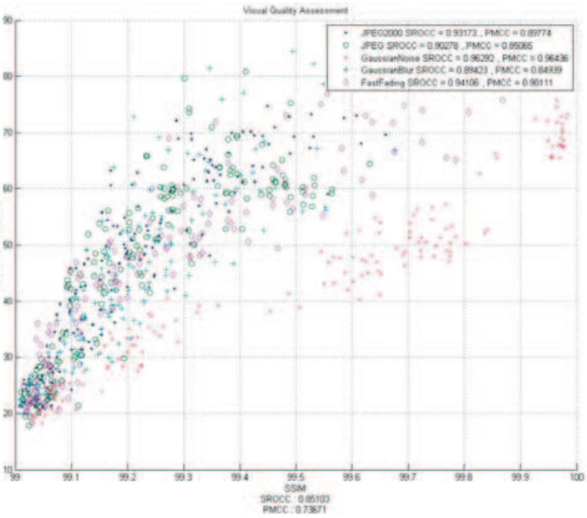

Figure 3. SSIM versus MOS for each degradation type

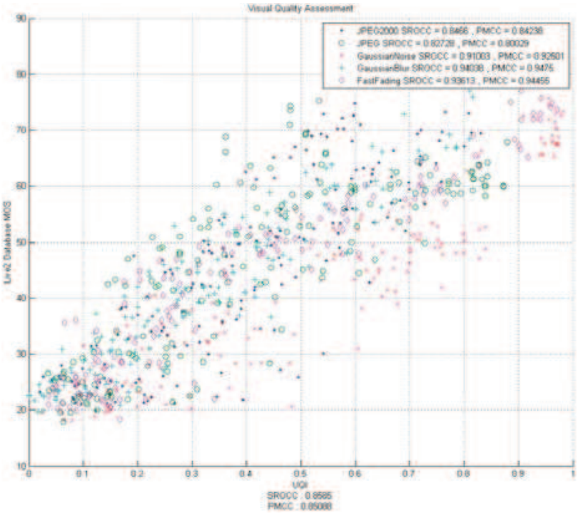

Figure 4. UQI versus MOS for each degradation type

The other two metrics are calculated somewhat different and address the structural dissimilarity of the input images rather than the squared difference of pixel values. Structural SIMilarity (SSIM) index between signals x and y is defined as:

$$SSIM(x,y)=[l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \qquad (3)$$

where α,β,γ are positive constants adjusting the relative importance of each of the three similarity components:

$$l(x,y)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1} \qquad (4)$$

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2} \qquad (5)$$

$$s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \qquad (6)$$

where *l* is a luminance comparison, *c* is contrast comparison and *s* is structure comparison. μ_x,μ_y are the averages of the signals within some window and σ_x,σ_y are the standard deviations. σ_{xy} is the normalized cross-correlation of the two signals in the same window. The authors of this metric propose the usage of a Gaussian window for measuring the similarity of each pixel and average the resulting SSIM of each pixel into a final vote. Our approach uses the Matlab implementation of this metric as presented in [9]. We tested the performance of the SSIM in the same way as for the PSNR and the resulting values and correlation coefficients are presented on Figure 3.

The UQI defined in [6] corresponds to the special case in SSIM when C1 = C2 = 0, which produces unstable results when either $(\mu_x^2+\mu_y^2)$ or $(\sigma_x^2+\sigma_y^2)$ is very close to zero. Although similar in implementation the UQI metric has different correlation to the MOS for different degradation types which can be seen in Figure 4. The results are obtained using the Matlab implementation given by the authors. The SROCC and PMCC values of each metric are in the range of 0.7~0.85 which indicates a high positive correlation between the metric and the human opinions. However, from the plots we can observe that the metrics perform very well under certain degradation types and far worse for others. This results in the fuzziness of the data points. A perfect metric would have a perfect correlation of 1.0 with the MOS and place the data points along the line *y*=*x*. In the next section we try increase the correlation of the metrics by fusing them using machine learning algorithm. The result should be a batter fit of the new metric to MOS which would linearize the data points and thus predict the perceived visual quality of each image with higher precision.

IV. Metric Fusion Using Regression

For the purpose of fusing the three metrics we use the C/C++ implementation of the ε-Support Vector Regression (ε-SVR) from LIBSVM [8]. In ε-SV regression, the goal is to find a function *f(x)* that has at most ε deviation from the actually obtained targets (the MOS) for all the training data (the metrics) and at the same time as flat as possible. The motivation to use ε-SVR is because it offers very vast training times and high flexibility for tuning the model.

We use 779 pairs of images *(X,Y)* from the LIVE2 database where *X* is the original image or the ground truth and *Y* is the respective distorted image. For each image pair we compute the three metrics (PSNR, UQI and SSIM) and create a pair of feature vector x_i and the MOS for that image y_i. This data matrix is then used to train a regression model. For better generalization of our model we use 90% of the images for training and 10% for testing. It should be noted that none of the test images are present in the training set and the training and testing sets contain images with all of the degradation types. We create 20 different permutations of the dataset and train 20 different regression models. Models are trained using

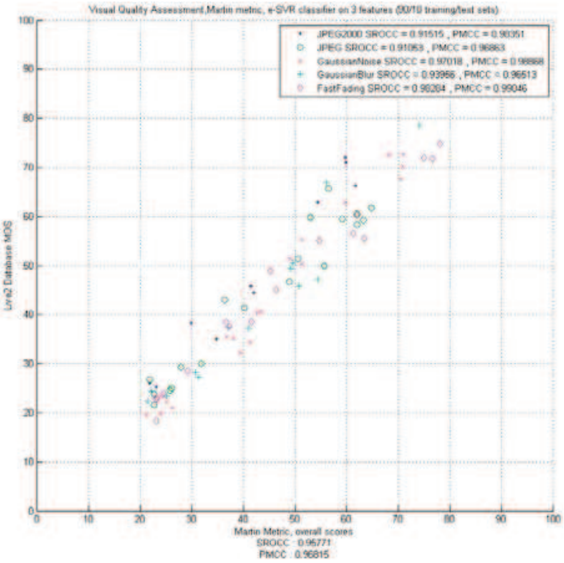

Figure 5. Results from the best ε-SVR model against the MOS. (note; there are only 10% of the data points used for testing)

3-fold cross-validation of the training sets for finding the optimal model parameters. We use a radial basis kernel mapping function for training and find three optimal parameters (cost, γ and ε). The results presented in Table 1 are the average values from these 20 models and the plot in Figure 5 are the data points from the best model. The regression model by its nature has the best and flattest curve fitting the data so as expected its the correlation coefficient PMCC is far better than any of the separate metrics. SROCC is also higher for the proposed metric. The data in Table 1 clearly shows the improvement of the proposed metric and the difference between the UQI and SSIM noted in section III.

The flexibility of the ε-SVR model comes from the selection of the ε parameter which controls the largest acceptable deviation of the output values from the training data. However choosing different values for ε apart from reducing the regression error makes the model more complex. Because of the nature of support vector machines this is achieved by adding more training feature vectors to the model as support vector. The increased number of support vectors in turn reduces classification time drastically. We performed a small test on one of our regression models by tweaking the ε parameter alone. Figure 6 shows the dependency of the number of support vectors and the overall accuracy of the model when the parameter ε is changed from infinity to zero. Our approach automatically chooses the ε value for which the model has the highest accuracy which has showed that produces many unnecessary support vector and isn't optimal.

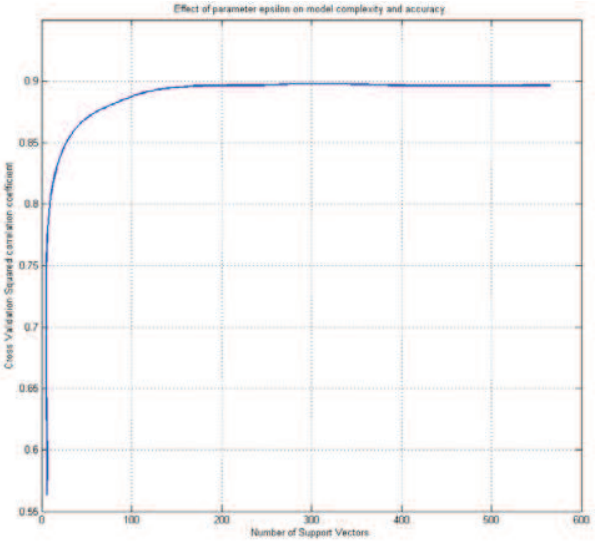

Figure 6. Effect from tweaking the parameter ε in our best regression model. From left to right the parameter is reduced from infinity to zero in 2^n steps.

V. Conclusion

In this paper we address the problem of full-reference image quality assessment by analyzing the performances of some of the most used metrics. We realize that any statistical measure or HVS based metric has its limitations when compared to actual human opinion because of the many approximations made for improving their speed and reducing the complexity. Therefore we present a novel approach for combining the PSNR, UQI and SSIM metrics with a machine learning method. By using epsilon support vector regression we fuse the votes from each metric into a final vote which shows superior correlation to mean opinion scores. When tweaked correctly the model also has a relatively low complexity making the regression very fast to compute, however this parameter tweaking requires exhaustive knowledge of the datasets and the used kernel function which is an unwanted factor. In the future we will be looking into further optimizing the parameter selection process as well as incorporating more sophisticated metrics into the regression model.

SROCC

Metric	JP2K	JPEG	Noise	Blur	FF	All
PSNR	0.889	0.84	**0.985**	0.781	0.89	0.819
UQI	0.846	0.827	0.91	**0.94**	0.936	0.858
SSIM	**0.931**	0.902	0.962	0.894	0.941	0.851
Proposed	0.915	**0.91**	0.97	0.939	**0.982**	**0.957**

PMCC

Metric	JP2K	JPEG	Noise	Blur	FF	All
PSNR	0.873	0.841	0.979	0.774	0.752	0.799
UQI	0.842	0.8	0.925	0.947	0.944	0.85
SSIM	0.897	0.85	0.964	0.849	0.901	0.736
Proposed	**0.983**	**0.968**	**0.988**	**0.965**	**0.99**	**0.968**

Table 1. SROCC and PMCC to the MOS for each metric on each degradation type

REFERENCES

[1] A. M. Eskicioglu, "Application of multidimensional quality measures to reconstructed medical images," *Opt. Eng.* 35(3), 778–785 (1996).

[2] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.* 43(12), 2959–2965 (1995).

[3] H. de Ridder, "Minkowsky metrics as a combination rule for digital image coding impairments," in *Human Vision, Visual Processing, and Digital Display III*, *Proc. SPIE* 1666, 17–27 (1992).

[4] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision* (A. B. Watson, ed.), pp. 207–220, the MIT press, 1993.

[5] Zhou Wang, Alan C. Bovik,Hamid R. Sheikh,Eero P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 13, NO. 4, APRIL 2004

[6] Zhou Wang, A. C. Bovik ," A universal image quality index," in Signal Processing Letters, IEEE, Vol. 9, No. 3. (2002), pp. 81-84.

[7] H.R. Sheikh, Z.Wang, L. Cormack and A.C. Bovik, "LIVE Image Quality Assessment Database Release 2", http://live.ece.utexas.edu/research/quality.

[8] C.-C. Chang and C.-J. Lin, \LIBSVM: a library for support vector machines." Sofware Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity" IEEE Transactios on Image Processing, vol. 13, no. 1, Jan. 2004.