doi:10.2498/cit.1001822

Selecting Low-level Features for Image Quality Assessment by Statistical Methods

Atidel Lahouhou, Emmanuel Viennet and Azeddine Beghdadi

Laboratoire de Traitement et Transport de l'Information, Institut Galilée – Université Paris 13, France

Image quality assessment is an important component in every image processing system where the last link of the chain is the human observer. This domain is of increasing interest, in particular in the context of image compression where coding scheme optimization is based on the distortion measure. Many objective image quality measures have been proposed in the literature and validated by comparing them to the Mean Opinion Score (MOS). We propose in this paper an empirical study of several indicators and show how one can improve the performances by combining them. We learn a regularized regression model and apply variable selection techniques to automatically find the most relevant indicators. Our technique enhances the state of the art results on two publicly available databases.

Keywords: image quality assessment, perceptual quality, JPEG, JPEG2000, structural similarity measure (SSIM), variable selection

1. Introduction

Considering the subjective appreciation of image quality in the image storage or transmission devices is very important. However, our knowledge about human perception mechanisms is still very limited. The few existing models of the Human Visual System (HVS) are established under very restrictive conditions. Since image quality is subjective in nature, its evaluation based on subjective experiments is a widely accepted solution. However, a lot of applications would gain from automatic real time image quality estimation (e.g. online QoS control in data networks used for video transmission).

Historically, objective image quality assessment methods were mainly based on simple mathematical measures such as the Euclidian distance between the pixels of the original image taken as the reference and its distorted version. The Peak Signal to Noise Ratio (PSNR) has been one of the most widely used metrics until now due to its analytical and computational simplicity. This makes the PSNR practical for the optimization of image coding, filtering and quality enhancement systems. But simple quantitative measures like PSNR or mean square error (MSE) do not always reflect the image distorsions as perceived by the HVS: for instance, two images with a large MSE distance can be considered nearly identical by the human observer.

In the last decade, numerous methods for image distortion evaluation inspired from the findings on Human Visual System mechanisms [2] have been proposed. For some known distortions, it is possible to develop a measure which exploits the a priori knowledge on the image degradation. These approaches focus particularly on the contrast sensitivity functions, on the perceptual decomposition into multiple channels, on the visual masking and on the visual attention. However, the resulting models are very limited in practice and function only in some simple and particular situations.

The HVS is able to quickly appreciate the quality of an image, even if its original version is absent, which suggests that it is probably based on a high level interpretation of the image, using a lot of knowledge about the scene at hand.

In this paper, we explore a new approach: trying to combine some widely used indicators to build a robust model estimating directly the subjective quality of the image. The paper is organized as follows: in the next section we discuss the problematic of image quality and introduce the concept of Mean Opinion Score (MOS). Then we present the image features we will use. Section 4 briefly presents the statistical model used and the variable selection technique. Finally, we discuss experimental results on several public images databases and suggest some directions for future work.

2. Image Quality

After years of research on image quality assessment, no definition of the concept of "quality" is universally accepted. In fact, the precise definition of image quality depends on the kind of images and on the application (still images or video, usage, ...) [8]. Obviously, the quality criteria should be different for machine vision applications and for image and video destined to human observation. In this work, we focus on applications where the final destination is the human visual system.

2.1. Image Quality Assessment Models

The goal of objective image quality assessment models is to automatically estimate the perceptual quality of images, in a way correlated with the human appreciation.

We distinguish three families of models in the literature:

- Full reference models, which use the original version of the image for the quality assessment of the processed version. The task reduces to a comparison of two images (fidelity). This comparison should be fast (easy to compute in real time) and correlated with human subjective appreciation. The vast majority of the proposed methods, including the ones proposed in this paper, fall in this category.
- Reduced reference models: in some applications (e.g. video-transmission), one can transmit along with the compressed image a feature vector giving relevant information to control the quality of the result image. Methods based on these features are fast, but their relatively poor performances restrict their use to some specific applications.
- No reference models: also called "blind models", they attempt to evaluate the quality of an image without access to its reference. This is a complex task, which requires prior information on the distorsion, the domain and on the interpretation of the scene.

2.2. Objective Versus Subjective Quality

Subjective image quality assessment is purely experimental. It consists of inviting a group of subjects to judge the quality of a set of images under well defined conditions, for instance, the protocols normalized by the International Telecommunication Union (ITU) [1].

The ITU test is divided into several sessions, each of 31 minutes and constituted of sets of at least 15 observers. Each image is shown to an observer (either the pair original/degraded, or only the degraded version) who is asked to score the image on a scale from 1 to 5 (see Table 1).

1	2	3	4	5
Very poor quality	Poor quality	Good quality	Very good quality	Excellent quality

Table 1. Mean Opinion Score classes.

It is known that subjective image quality varies from one individual to another: usually, the scores given by different individuals are not identical. The oberver's score depends on his general experience (if he/she is expert in image processing or not), on his personal appreciation and may vary according to his mood. To alleviate this problem, an average score is computed over all observers. This *Mean Opinion Score* is denoted by MOS.

Another approach attempts to overcome those drawbacks by developing objective image quality assessment models that describe the influence of several physical features of the image [11]. These models still suffer from certain inconsistencies.

In this work, we propose a hybrid solution based on machine learning: we use a set of labeled images, for which the MOS has been recorded, to build a model able to generalize to novel unseen images.

3. Feature Extraction

In this section, we describe the images features used as input to the quality estimation model. These features are extracted from images pairs (original and degraded).

Several previous studies (e.g. [11]) concluded that the most important information for image quality assessment is carried by the luminance signal. Hence, all our images have been converted from the RGB space to the YCbCr space (where CbCr are the chrominance components and Y is the luminance signal), and the features derived from Y only.

The table below lists the features chosen. They can be grouped in two categories: simple local statistical statistics (μ, σ, MSE, MAX-COVAR, MAX-MSE), and composite indices devised in the litterature to directly estimate the image quality (PSNR, SSIM, SNR_{WAV}). Lots of other features could be added, but the main objective of this work is to test and combine well-known indices and to get a fast and simple estimator.

PSNR is a classical index defined as the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. It is given by:

$$\text{PSNR} = 20\log_{10}\frac{\max(I)}{\sqrt{\text{MSE}}}$$

where $\max(I)$ is the maximal possible value the image pixels can take, and MSE is the Euclidian distance between the original and the degraded images.

SSIM is an objective image quality measure proposed by Wang et al. [11], which compares two images, a reference image and its distorted version, using information about luminance, contrast and “structure”. The SSIM between two images x and y is thus based on pixels means and standard deviations:

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

C_1 and C_2 are positive constants chosen empirically to avoid the unstability of the measure. Note that all SSIM components ($\mu_.$, $\sigma_.$) are also considered invidually in our experiments, allowing us to validate the combination proposed by SSIM.

Feature	Definition
μ_x, μ_y	Mean pixels values of the original and processed images, respectively.
σ_x, σ_y	Standard deviation between the original and processed images pixels, respectively.
σ_{xy}	Covariance between pixels values of the original and processed images.
MSE	Mean square error between original and processed images.
MAX-COVAR	Maximal covariance between 8x8 blocks of the two images
MAX-MSE	Maximal mean square error between 8x8 blocks of the two images
PSNR	Peak Signal to Noise Ratio.
SSIM	*Structural Similarity Measure*.
SNR_{WAV1}	Wavelet-based distortion measure using biorthogonal 9/7 wavelets.
SNR_{WAV2}	Wavelet-based distortion measure using cubic spline wavelets.

Table 2. Images features extracted.

SNR_{WAV} is another image distortion measure [3], based on wavelet decomposition. The multiresolution analysis computed by the wavelet transform allows to take into account the effect of the distorsions at different scales. The measure is defined as $\text{SNR}_{\text{WAV}} =$

$$20\log_{10}\left(\frac{\sum_{k,l,d}\max_j 2^{-jsp}|c_j^d(k_j,l_j)|^p}{\sum\max_j 2^{-jsp}|c_j^d(k_j,l_j)-\hat{c}_j^d(k_j,l_j)|^p}\right)^{1/p}$$

where $c_j^d(k_j,l_j)$ are the wavelet coefficients of the distorted image and s and p are positive constants (Besov parameters). This definition simply indicates that we consider (like in PSNR) the maximum of the absolute value of the wavelet coefficients over the trees spreading all the resolution levels and corresponding to the same spatial location and orientation. Following [3], we used two families of wavelets, the biorthogonal 9/7 wavelets and the cubic spline wavelets giving the measures SNR_{WAV1} and SNR_{WAV2} respectively.

4. Estimation Model and Variable Selection

A lot of statistical models can be used to build an estimator of the MOS based on the features described above. We chose to stick with a simple and efficient approach, a regularized multi-dimensional polynomial estimator (of order 1 or 2), implemented by KXEN K2C/K2R components[1]. This model relies on a *Structural Risk Minimization* approach to optimize the parameters and hyperparameters (encoding of the variables and ridge regression) [5]. These parameters are estimated on a set of labeled images (with known MOS), and then the model can be applied to new images. It is important to understand that the variables are encoded using a non-linear (stepwise) procedure before being used by the polynomial regressor.

This kind of statistical modeling can supply an accurate estimation of the target variable (MOS) (Figure 1), and can also estimate the contribution of the various features (for a detailed study of features selection techniques, see [6]). Basically, the features are ranked according to their weight in the polynomial expression. This allows to take into account eventual correlations between the features and cases where individual features are not correlated to the target variable, but their (non linear) combination carries valuable information.

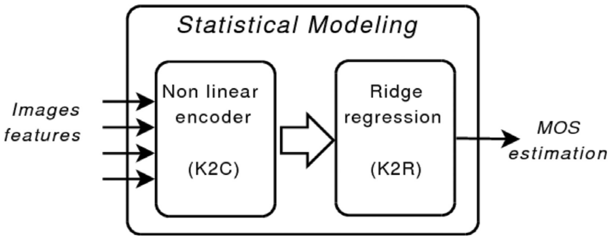


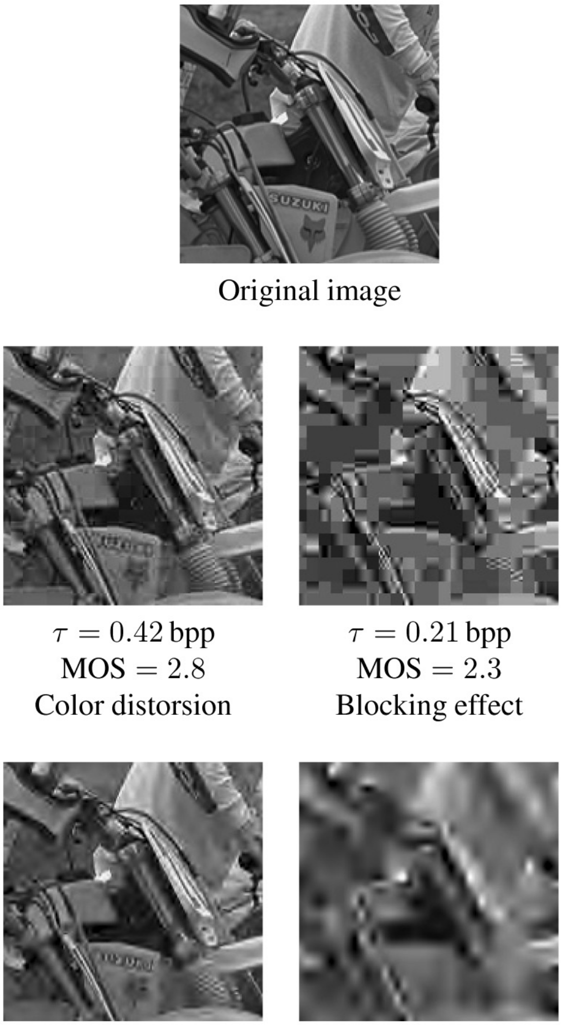
Figure 1. Modeling of the MOS.

5. Experimental Results

5.1. Image Databases

We used two distinct publicly available databases, composed of natural images, original and compressed, using the widely used JEPG and JPEG 2000 algorithms. Degradations caused by JPEG coding are: blocking and ringing effects, blur and color distortion (Figure 2).

The LIVE database [10] contains 29 high resolution color images (24 bits/pixel) of different sizes (typically 754x640). These images have been encoded at different bit rates (from 0.028 bit-per-pixel to 3.34 bpp) using JPEG and JPEG2000 algorithms, generating 460 distinct images. The induced distortion levels cover a large range of quality: from excellent quality where artefacts are not visible, to very poor quality where distortions are annoying (see Fig-



Original image

$\tau = 0.42\,\text{bpp}$
MOS $= 2.8$
Color distorsion

$\tau = 0.21\,\text{bpp}$
MOS $= 2.3$
Blocking effect

$\tau = 0.13\,\text{bpp}$
MOS $= 2.2$
Ringing effect

$\tau = 0.04\,\text{bpp}$
MOS $= 1.9$
Blurring

Figure 2. JPEG (above) and JPEG2000 (below) artefacts for different compression levels τ, with associated MOS.

[1] http://kxen.com

ure 2). For each image, the Mean Opinion Score (MOS) has been estimated by experimentation under specified conditions recommended by the International Union of Telecommunications as detailed in [1].

The IVC database has been published by the IVC (*Image Video and Communication Lab*, `http://www2.irccyn.ec-nantes.fr/ivcdb/`), University of Nantes [4]. This database contains 10 original images that were subjected to JPEG, JPEG2000 and blurring image processing algorithms to generate 170 processed images. Subjective evaluations of images were carried out by 15 observers using double stimulus method mentioned above. The subjective quality scores (MOS) were derived from the obtained quality scores which are in the range $[0, 1]$ by using a psychometric function. This methodology has been approved and recommended by the Video Quality Experts Group (VQEG) [9]. To fairly compare results on both databases, we selected only the JPEG and JPEG2000 images of the IVC database.

Features	**LIVE**	**IVC**	**L/IVC**
μ_x, μ_y, σ_x, σ_y, σ_{xy}	0.76	0.35	0.33
PSNR	0.90	0.69	0.42
SSIM	0.94	0.76	0.76
$\mathrm{SNR}_{\mathrm{WAV1}}$	0.94	0.78	0.78
$\mathrm{SNR}_{\mathrm{WAV2}}$	0.94	0.75	0.77
All features	0.95	0.90	0.43
SSIM, PSNR, MSE	0.95	0.92	0.80

Table 3. MOScorrelations on three databases for different variables, linear regression models (10 fold cross-validation, test set, average on ten runs).

Features	**LIVE**	**IVC**	**L/IVC**
μ_x, μ_y, σ_x, σ_y, σ_{xy}	0.88	0.80	0.44
All features	0.95	0.88	0.72
SSIM, PSNR, MSE	0.95	0.91	0.79

Table 4. Same results with second order regression models.

5.2. Experimental Setup

All presented results are obtained by a ten-fold cross-validation procedure [7]: the image set is splitted in three parts, the estimation set with 80% of images, the validation set with 10% of the images, and a test set with the remaining 10%. On each of the ten runs, the models parameters are fitted on the estimation set, and generalization (hyper-parameters) is controlled by observing the error on the validation set. Finally, the model is applied on the test set.

The performance measure is simply the linear correlation rate between the estimated MOS and the real MOS, averaged on ten distinct test sets.

We present (Tables 3 and 4) the results on the two databases (LIVE and IVC), obtained with different feature sets. The right column, L/IVC, corresponds to experiments where the models are fitted on LIVE images, but applied (tested) on the IVC images.

5.3. Discussion

Both used databases were designed to test JPEG /JPEG2000 image quality assessment methods, but they are quite different. In fact, the models performances are significantly lower on IVC, suggesting that the subjective evaluation is harder to reproduce on these images. The first line of the two tables gives the correlation rate of the models based only on the features (μ_x, μ_y, σ_x, σ_y, σ_{xy}). It is surprising to note that these models perform so badly on IVC, while they give quite good results on LIVE. The SSIM index combines these same five features. We tried to combine these features with polynomials of order 1 or 2 (Table 4). Although the latter performs better, it is worth noting that the ad-hoc non linear combination computed by SSIM [11] gives superior results: SSIM incorporates a priori expert knowledge which cannot be learned using only a few images.

The models built using all 12 features, denoted by "*All features*" in the tables, consistently lead to better results on LIVE and IVC. The gain is especially significant on IVC (we found that all models except the first one give approximately identical results on LIVE database). This is an important result, as it shows that we can enhance the widely used SSIM index by combining it with simple features.

In order to build faster models, we applied a variable selection procedure to determine the most important features. The ranking of the variables is presented in Figure 3. Three features (SSIM, PSNR and MSE) contribute most. We then built models using only these three features. Reducing the number of features also enhances the robustness of the models, as some features may only add noise. The resulting model (bottom lines) is effectively the best one. In particular, it performs very well on the L/IVC experiment, where we estimate the models on LIVE images, but apply them (and measure correlation) on IVC. This is the harder task, stressing the generalization ability of the models.

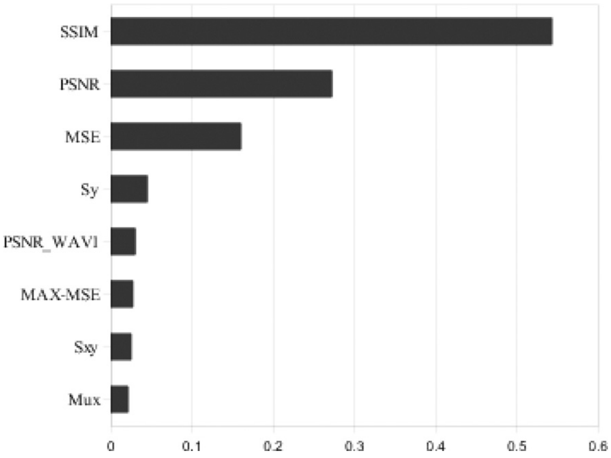


Figure 3. Variables contributions.

Finally, let's note that, to our disappointment, the SNR$_{\text{WAV}}$, while beeing more complex to compute and performing quite well individually (it is roughly equivalent to SSIM), does not enhance the correlation rate when combined with other features. This is a surprise, because we thought that it would carry information of different nature. A possible explanation is that JPEG artefacts are well detected using only local features and do not require multiresolution analysis.

This work is a preliminar step in the development of new automatic image quality assessment methods. We are now analyzing the errors (images where the MOS discrepancy is higher) and trying to propose new feature extractors to handle these cases. For instance, one could use a convolutionnal neural network to process the images and extract relevant information. The methodology proposed in this paper allows to quickly test new ideas and determine the best combination of features.

Another interesting extension of our work will be to devise an estimator based only on the transmitted image, without access to the reference.

References

[1] ITU-R RECOMMENDATION BT.500-7: Methodology for the subjective assessment of the quality of television pictures, 1995.

[2] Special issue of Signal Processing on Image quality assessment, vol. 70, 1998.

[3] A. BEGHDADI, B. PESQUET-POPESCU, A new image distortion measure based on wavelet decomposition. In *Proceedings of the 6th International Symposium on Signal Processing and Its Applications (ISSPA '03)*, volume 1, pages 485–488, Paris, July 2003.

[4] M. CARNEC, P. LE CALLET, D. BARBA, Objective quality assessment of color images based on a generic perceptual reduced reference. *Image Communication*, 23(4):239–256, 2008. ISSN 0923-5965.

[5] F. FOGELMAN-SOULIÉ, E. MARCADÉ, L'industrialisation des analyses – besoins, outils & applications. *MODULAD*, 1(38): 140–158, July 2008.

[6] I. GUYON, S. GUNN, M. NIKRAVESH, L. A. ZADEH, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 3540354875.

[7] T. HASTIE, R. TIBSHIRANI, J. FRIEDMAN, *The Elements of Statistical Learning*. Springer, 2001.

[8] T. J. W. M. JANSSEN, F. J. BLOMMAERT, A computational approach to image quality. *Displays*, 21(4): 129–142, 2000.

[9] A. NINASSI, P. L. CALLET, F. AUTRUSSEAU, Pseudo no reference image quality metric using perceptual data hiding. In *SPIE Human Vision and Electronic Imaging*, volume 6057, pages 146–157, 2006.

[10] H. SHEIKH, Z. WANG, L. CORMACK, A. BOVIK, LIVE image quality assessment database release 2. `http://live.ece.utexas.edu/research/quality`.

[11] Z. WANG, A. C. BOVIK, *Modern Image Quality Assessment*. Morgan and Claypool Publishing Company, New York, 2006.

Received: June, 2009
Accepted: April, 2010

Contact address:
Emmanuel Viennet
Laboratoire de Traitement et Transport de l'Information
Institut Galilée
Université Paris 13
e-mail: `emmanuel.viennet@univ-paris13.fr`

ATIDEL LAHOUHOU received the Ingénieur d'Etat degree in computer science from the University of Constantine (Algeria) in 2000, the Magister degree from the University of Jijel (Algeria) in 2003. She is currently preparing her PhD degree jointly at Ecole Nationale Polytechnique of Algiers (Algeria) and Institut Galilée of the University Paris 13 (Paris). Her research work concerns image quality assessment and neural approaches.

DR. EMMANUEL VIENNET is Professor at the University Paris 13 (Institut Galilée). He received his PhD in Computer Sciences from the University Paris XI in June 1993. His research activities are in the field of statistical pattern recognition, applying tools like artificial neural networks and support vector machines to data mining or image processing tasks. His recent work concerns image quality assessment and social network analysis.

DR. AZEDDINE BEGHDADI is Professor at the University of Paris 13 (Institut Galilée) and a researcher at L2TI laboratory where he does all his research in image and video processing. He received his PhD in Physics (specifically: optics and signal processing) from the University Paris 6 in June 1986. He has published over 160 international refereed scientific papers. He is a funding member of the L2TI laboratory. His research interests include image quality enhancement and assessment, compression, bio-inspired models for image analysis and physics-based image analysis. Dr. Beghdadi has served as conference chair and as session organizer and is a member of the organizing and technical committees for many IEEE conferences. Dr Beghdadi is a Senior member of IEEE.