Correlating Objective and Subjective Color Image Quality Evaluation for JPEG 2000-Compressed Images

Adrian STOICA, Constantin VERTAN, Christine FERNANDEZ-MALOIGNE

IRCOM-SIC Laboratory, UMR-CNRS 6615
University of Poitiers
Bat. SP2MI, Téléport 2, BP 30179, 86962 – FUTUROSCOPE Chasseneuil Cedex, France

{stoica, vertan, maloigne}@sic.sp2mi.univ-poitiers.fr

Abstract

The subjective evaluation of the color image quality came into focus because of the "failure" of the objective criteria to offers good results in terms of human perception. In general, objective and subjective evaluation of compressed images can lead to significant different conclusions. We show that the classical MSE measure can become correlated with perceptual experiments for an inter-color distance based on a plane, triangular color plot.

Keywords: color image quality, color model, color distance

1. Introduction

In recent years there has been a growing number of digital image and video application, which all imply displaying, storing and transmitting large volumes of image data. Obviously, image data compression is a must and various lossless or lossy compression have been developed, notably the JPEG-family standards.

The evaluation of the compression quality is a simple and straightforward task in the case of lossless compression: standard criteria (compression ratio, execution time, etc) can be used. In the case of lossy compression the main difficulty arises in describing the type and the amount of degradation induced in the reconstructed image. The need of evaluating image quality in a human-assessment-consistent way has led to several approaches in image quality evaluation. There are two major types of image quality criteria: objective criteria and subjective (human-judgement-based) criteria. Since subjective image quality measure exhibit some inherent drawbacks (the use of a normalized evaluation room, a large panel of human observers, etc), there has been a great deal of interest in developing quantitative measures, either in numerical or graphical form, that can be consistently used as a substitute [1], [3], [5].

The remainder of the paper is organized as follows: Section 2 presents the classical MSE objective quality measure, updated with a new inter-color distance; Section 3 presents some subjective quality evaluations procedures; Section 4 presents the technical characteristics of the evaluation experiments; Section 5 presents the experimental results of image quality evaluation with both objective and subjective methods for JPEG 2000-compressed color images, based on two largely available implementations: JJ2000 and Kakadu. The paper ends with some conclusions and ideas for further work.

2. Objective Quality Measures

The classical criteria are derived from the idea of measuring the closeness between two digital images by exploiting the statistical differences in the pixel values. The most known objective image quality measures are the Peak Signal-to-Noise Ratio (PSNR) and the Mean Squared Error (MSE), which are widely criticized for not correlating with the perceived image quality, especially for color images.

The Mean Squared Error (MSE) between the original and the decompressed images is defined as follows:

$$MSE = \frac{1}{MN} \sum_{j=1}^{M} \sum_{k=1}^{N} \left\| F(j,k) - \tilde{F}(j,k) \right\|^2 \qquad (1)$$

where $F(j,k)$ is the original image, $\tilde{F}(j,k)$ is the decompressed image, M and N are the height and the width of the image.

If the MSE decreases to zero, the pixel-by-pixel matching of the images becomes perfect. If the MSE is small enough, this will presumably correspond to a high quality decompressed image. For a given image and a given compression algorithm the value of MSE increases as the compression ratio increases. However, the MSE can increase in an arbitrary manner and does not follow the fact that the image quality is always decreasing. The same MSE does not guarantee that two images are of the same subjective quality.

Although the MSE is widely used, it is also known to be not well-correlated with the human judgements. The MSE does not take into account either the correlation between color components, or the neighborhood of image pixels (A subjective study showed that the blocking effect distortion is ten times more objectionable than equal energy white noise [6]. In other words, humans are more sensitive to the structure of the noise and not just the noise energy.)

Basically, the MSE measures the average distance between colors within the two compared images. The classical inter-color distances are based on standard colorimetric representations, being metrics (or almost metrics) in the color gamut space (a subset of $\mathbf{R}^3$). We propose to investigate a new approach, inspired by the reduced ordering principle of Barnett [9] and also used in multivariate data visualization [11] : namely, the colors (multivariate data vectors) are replaced by scalars, or other "familiar" two-dimensional objects, that can be grouped, compared, and plotted with more ease (more computationally attractive solutions) and are more suited for human perception. Examples of such mappings are the Chernoff faces [10], [11], the Andrews curves [8], the basic parallel coordinates [11] and the star glyphs [11].

We will further explain the use of star glyphs for representing colors in the two-dimensional (black-and-white colored) plane. A star glyph associated to a p-dimensional ($p>2$) data vector is a plane polygon obtained by connecting p points in the two-dimensional plane. The points are taken along p origin-concurrent, equally-spaced axes that span the two-dimensional plane (i.e. the angle between successive axes is $2\pi/p$). Each axis is associated to one component of the vector data, which is plotted according to its magnitude (see figure 1). Thus, each color is mapped into a plane triangle and comparing colors reduces to comparing their corresponding triangles. Obviously, the triangle's shape is related to the nature of the color (properties like hue, saturation, colorfulness, etc.) and the overall size of the triangle relates to the luminance or brightness of the color.

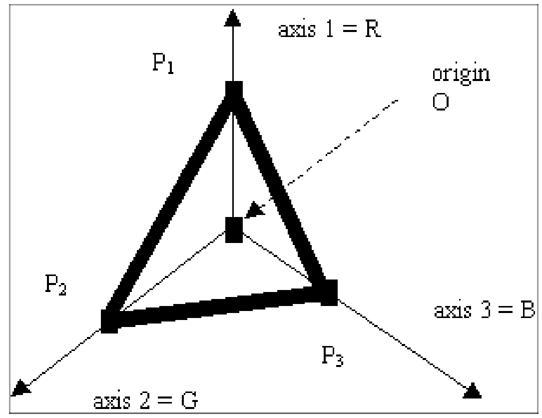

***Figure 1**. Star glyph associated to a three-component color vector **C** = (R, G, B). The angles between axes are 2π/ 3 = 120°. The star glyph is a triangle (black line).*

It can be easily accepted that the similarity measure between the colors is measured by the area of the common part (intersection) of the two associated triangles. In order to construct a symmetrical and normalized measure we will normalize to the maximum area of the two triangles. Thus, the similarity between colors $\mathbf{C}_1$ and $\mathbf{C}_2$, represented by their corresponding triangles $T_{\mathbf{C}1}$ and $T_{\mathbf{C}2}$ is:

$$Sim(\mathbf{C}_1,\mathbf{C}_2) = \frac{Area(T_{\mathbf{C}_1} \cap T_{\mathbf{C}_2})}{max(Area(T_{\mathbf{C}_1}), Area(T_{\mathbf{C}_2}))}. \qquad (2)$$

This can be transformed into an inter-color distance according to the following equation:

$Dist(C_1, C_2)=1-Sim(C_1, C_2)$. (3)

3. Subjective Quality Evaluation

The problem that there does not exist an objective image quality measure that perfectly reflects the subjective impression of a human observer is well-known. The various objective image quality measures indicate only roughly the image quality and are not properly defined for color images.

Subjective tests provide the foundations for building vision models. At the same time, they are the only true benchmark for evaluating the performance of the various perception-based image processing tools. Unfortunately, perceptual responses cannot be represented by exact figures; due to their inherent subjectivity, they can only be statistically described. Even in psychophysical threshold experiments, where the task of the observer is just to give a yes/no answer, there exists a significant variation between observers. In the evaluation of supra-threshold artifacts, these differences become even more pronounced, because the objectionability of artifacts strongly relates to the observers expectations and presumptions, as to the intended application. Previous observer experiences also lead to a different weighting of the artifacts [2], [7].

Subjective assessment of visual quality has been normalized in the ITU-R Rec.500 [4], which suggests standard viewing conditions, criteria for the selection of observers and test material, assessment procedures, as well as data analysis methods. While targeted at the subjective assessment of television pictures, most of it directly applies to still images as well. In particular, it describes the Double Stimulus Continuous Quality Scale (DSCQS) and the Double Stimulus Impairment Scale (DSIS).

In a DSCQS test, viewers are shown stimulus pairs consisting of a “reference” and a “test” stimulus, which are presented twice in an alternating manner, with the order of the two chosen randomly for each trial. Subjects are not informed which the reference is and which is the test stimulus. They rate each of the two pairs separately on a continuous quality scale ranging from “bad” to “excellent”. Analysis is based on the difference in rating for each pair, which is calculated from an equivalent numerical scale from 0 to 100. The DSCQS has been shown to work reliably even when the quality of test and reference stimuli are rather similar, because it is quite sensitive to small differences in quality.

In a DSIS test, the reference is always displayed before the test stimulus, and both are shown only once. Subjects rate the amount of impairment in the test stimulus according to a discrete five-level scale ranging from “very annoying” to “imperceptible”. DSIS is the referred method when evaluating clearly visible impairments.

4. Experiments

Both objective and subjective tests were performed on 12 images from the Kodak database (768x512 pixels, 24 bpp RGB color images). Each image was compressed at various bit-rates, ranging from 4 bpp to 0.062 bpp (4 bpp, 3 bpp, 2 bpp, 1 bpp, 0.5 bpp, 0.375 bpp, 0.25 bpp, 0.187 bpp, 0.125 bpp, 0.062 bpp) and decompressed according to the JPEG 2000 standard. Two JPEG 2000 compression algorithms were used, namely the Java implementation JJ2000 (version 5.1) and the C++ implementation Kakadu (version 3.4).

The objective quality is evaluated by the MSE, using the classical L2 (1) and triangular-model-based (2) inter-color distances. In order to visualize the variation of the two measures over the entire range of compression rates, a MDS (Multidimensional scaling) technique was used. Multidimensional scaling is a well-known statistical tool, intensively used in psychology and other cognitive sciences. [12] Basically, the MDS allows to compute a numerical representation for any arbitrary objects, characterized only by their pairwise interactions. Usually, such inter-object interaction is a numerical dissimilarity measure (or difference, or distance) between pairs of objects [13], [14]. The MDS allows 2D or 3D-layout plot of different objects, characterized by their inter-distances only. In our case, the objects are the images compressed at different compression rates, and the distances, are the MSE measures.

The subjective tests were performed with a panel of ten observers, with different image processing backgrounds, which were evaluated for the visual acuity and a normal vision of color using the Ishihara test. The observers were asked to evaluate the photographic samples of the degraded images, as displayed on a 24" Sony high-resolution computer monitor, with a Trinitron tube. The screen was calibrated using a Gretag Macbeth EYE-ONE monitor Mach 1.1 calibrator. The computer characteristics were: an AMD ATHLON XP 2100+ (1.73 GHz) processor, 1 Gb RAM, and a NVIDIA GeForce4 TI 4600 128Mo video card. This configuration allows a pixel to pixel image presentation on the screen, without image resizing or other internal interpolations. The experiments were made up in a special room, which was built according to the international standards ITU-R 500-10 and ISO 3664. The room walls were of neutral gray, in order to avoid the effect of flare. The room illumination was totally controlled variable artificial light, measured by a Minolta T-10M luminance-meter, and fixed at 62 lux. The viewing distance was set at 60 cm, in order to allow a very good perception of image details.

The evaluation technique was a hybrid method based on the DSCQS, in which the quality scale was replaced by a choice of one of the two compressed images. Three images were presented at the same time, as follows: in the center the original image, and on the left and right, the JJ2000 and Kakadu compressed images at the same bit-rate. The positioning of the two compressed images was randomly chosen, at each presentation, so the observer did not know which was the JJ2000 or the Kakadu compressed image. At each presentation the observer had to choose one of the two compressed images that he liked more, from a visual point of view.

5. Results

Figure 2 shows the 3D representation of the MDS plots obtained from the two MSE measurements for a Kodak test image, compressed with the JJ2000 compression algorithm, at all the compression bit rates. The plots of the MSE measure for the Kakadu-compressed image exhibit a similar, S-shape.

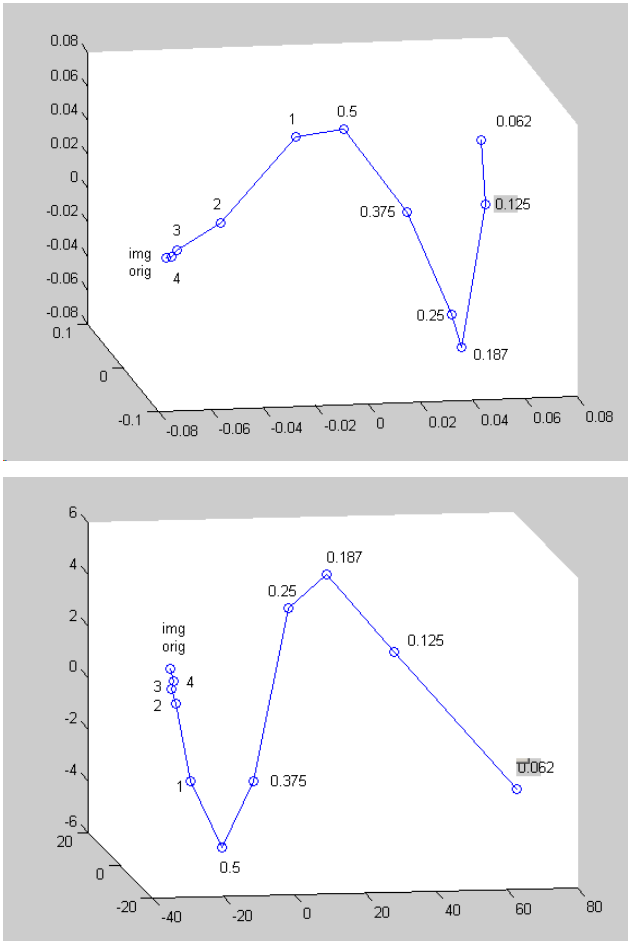

Figure 2*. MDS plots for the MSE quality measure of a Kodak image compressed by the JJ2000 algorithm: a) triangular-model-based inter-color distance b) L2 inter-color distance. The numbers associated to the data points correspond to the compression bit rates and the axes correspond to the MDS distances.*

The first analysis of the two MDS representations leads to the conclusion that the quality of the compression doesn't have the same behavior for all the compression bit rates ranges. We can easily identify three zones of similar behavior, separated by the inflexion points of the plotted curve, placed at about 0.5 bpp and 0.2 bpp.

According to the MDS plot of the L2-distance-based MSE, the image quality variation is not uniformly distributed with respect to the bit-rate. This aspect is not quite accurate, fact proved also by the subjective evaluations. For example, during a subjective quality evaluation experiment, one noticed that the image quality variation is not that significant between the images compressed at 0.375 bpp and 0.25 bpp, like the classical MSE may suggest. On the other hand, the MDS plot of the triangular-model-based-distance MSE, we can identify a better spatial distribution of different points within the bit rate rage. This means that the use of a measure which taking into account color-specific distances (like (2)) leads to an image quality evaluation which is closer to the subjective evaluations.

Another conclusion might be that the luminance invariance masks the color differences.

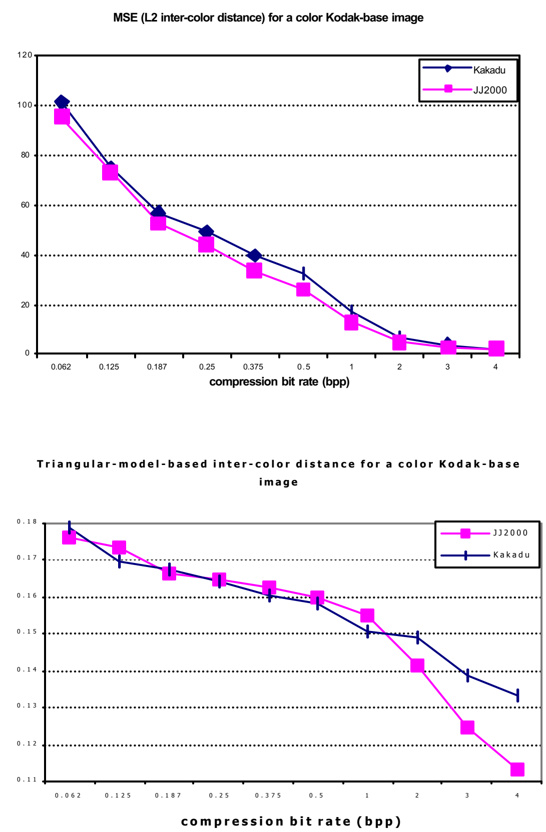

Figure 3*. Objective evaluation (MSE) of image quality vs. bit rate for a Kodak test image, compressed with two JPEG2000 algorithms: a) L2 inter-color distance; b) triangular-model-based inter-color distance.*

Figure 3 shows the MSE variation for a compressed image with respect to the obtained bit rate. We can still identify (although not very visible) two inflexion points, located at the bit rates of 0.5 bpp and 0.187 bpp, marking the same three intervals detected in the MDS plot. If the MSE based on the inter-color distance form (3) is used, the three intervals are characterized by different behaviors of the compression algorithms: the JJ2000 algorithm performs better for small compression bit rates (low compression), whereas the Kakadu algorithm performs better in the range of medium bit rates.

The triangular-model-based MSE was applied also to the classical JPEG-LS compression, for which the compression distortions are quite different from the ones encounter in the JPEG2000 compression. Figure 4 presents the L2-distance-based MSE and the triangular-model-based MSE variation for a compressed image with respect to the obtained bit rate, for the two JPEG2000 implementations and the classical JPEG-LS.

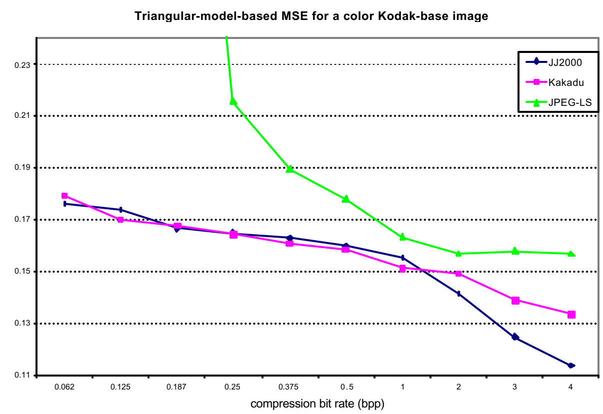

***Figure 4**. Objective evaluation (MSE) of image quality vs. bit rate for a Kodak test image, compressed with the classical JPEG-LS and two JPEG2000 implementations: a) L2 inter-color distance; b) triangular-model-based inter-color distance.*

In comparing the JPEG-LS with the JPEG2000 the proposed method one can easily observe a better performance of the new standard, with respect to the same three behavior intervals. According to the triangular-model-based MSE for high compression bit rates (high compression), over 0.24 bpp, the classical JPEG creates an image with very important distortions, especially from a color point of view. Although the number of color decreases significantly, false colors are created, as one can see in the Figure 6 *e)*. These results can be easily verified by a subjective image quality evaluation experiment.

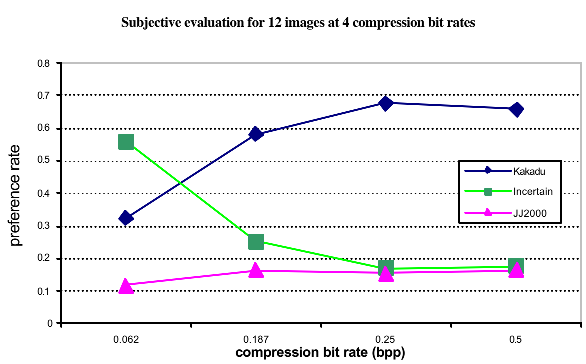

***Figure 5**. Quality preferences obtained from a subjective quality evaluation experiment for 12 Kodak test images at 4 compression bit rates.*

Figure 5 presents the results for the subjective quality evaluation of all 12 Kodak test images, at all compression bit rates. We can identify the same preferences in choosing the Kakadu algorithm for medium compressions, and the same indecision for the strong compression bit rates, as presented in the triangular-model-based MSE plot from figure 3 b). The subjective preferences are thus correlated to the proposed triangular-model-based MSE measurements. Moreover, the subjective results are in contradiction with the measurements obtained by the L2-distance MSE, according to which the JJ2000 compression algorithm performs better for all the compression bit rates (Figure 3 a)).

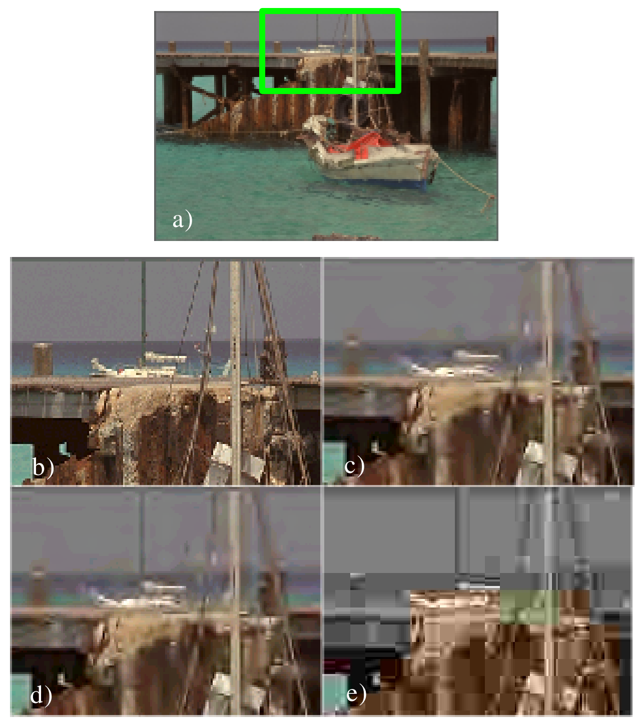

***Figure 6**. Sub-images (a) and (b) show the original image and a magnified sub-region. The sub-images (c), (d) and (e) presents the magnification of the compression result for JJ2000 (c), Kakadu (d) and JPEG-LS (e) for the subjective image quality evaluation.*

Figure 6 presents a particular subjective evaluation situation, from we can easily understand the choice made by the human observers. From a visual point of view the image compressed with the Kakadu algorithm (Figure 6 *d*) was preferred to the one compressed with the JJ2000, especially because of the missing structural elements in image *c)* (the JJ2000 compression). This choice is in total contradiction with the L2-distance MSE, for which the JJ2000 perform better for all the images and compression rates, and in correlation with the triangular-model-based MSE, which for this case gives a similar result as the subjective evaluation.

6. Conclusions

In this contribution we proposed the use of a new inter-color distance, based on the plane, triangular color representation. Although it does not attempt to solve the problem of pixel based measures (such as MSE) which do not take into account the spatial proprieties of colors, this inter-color distance is proved to exhibit (in the framework of compressed image quality evaluation) significant correlations with the perceptual evaluations. The measurements suggest also that a perceptual-coherent distinction is possible between different image compression codecs by simple MSE-type measures.

7. References

[1] S. Daly, "The Visible Differences Predictor: an Algorithm for the Assessment of Image Fidelity", in *Digital Images and Human Vision*, edited by A. B. Watson, MIT press, pp. 197-206, 1993.

[2] G. Deffner et al.: "Evaluation of display-image quality: Experts vs. non-experts". in *SID Symposium Digest*, vol.25, pp.475-478, 1994.

[3] M. Esckicioglu and P. Fisher, "Image Quality Measures and their performances", *IEEE Transactions on Communications*. Vol. COM-43, No. 12, pp. 2959-2965, Dec., 1995.

[4] ITU-R Recommendation BT.500-10: "Methodology for the subjective assessment of the quality of television pictures. ITU, Geneva, Switzerland, 2000

[5] P. Carrai, I. Heynderickx, P. Gastaldo, R. Zunino, "Image quality assessment by using neural networks", in IEEE International Symposium on Circuits and Systems, ISCAS 2002, Volume 5, pp. 253-256, 2002.

[6] M. Miyahara and K. Kotani, "Block Distortion in Orthogonal Transform Coding-Analyses, Minimisation and Distortion measure", *IEEE Trans. Commun*., Vol. COM-33, No. 1, pp. 90-96, Jan. 1985.

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

[8] Andrews D. F. "Plots of High-Dimensional Data", *Biometrics*, 28: 125-136, 1972.

[9] Barnett V. "The Ordering of Multivariate Data". *J. of the Royal Society of Statistics* A, 139 (3): 318-355, 1976.

[10] Chernoff H. "The Use of Faces to Represent Points in k-Dimensional Space Graphically". *J. of the American Statistical Association*, 68: 361-368, 1973.

[11] Krzanowski W. J. *Principles of Multivariate Analysis: A Users's Perspective*, Clarendon Press, Oxford, 1993.

[12] Steyvers M. *Encyclopedia of Cognitive Science*, ch. Multidimensional Scaling. Macmillan Reference Ltd., 2002

[13] Cox T.F. and Cox M.A.A. *Multidimensional Scaling*, Chapman & Hall, 1994.

[14] Kruskal J.B. and Wish M. *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA, 1978.

8. Biography

Adrian Stoica received his B. S. degree in Electronics and Telecommunications from the Polytechnic University of Bucharest in 2000 and a Master in Transmission of Information: Informatics, Image and Automatic from University of Poitiers in 2001. Actually he is a Ph.D. student at the University of Poitiers and Polytechnic University of Bucharest. His work focused on the image compression and image quality evaluation, with application to JPEG2000 image compression standard.