Objective Image Quality Measures of Degradation in Compressed Natural Images and their Comparison with Subjective Assessments

Alison K. Cheeseman[1,2], Ilona A. Kowalik-Urbaniak[1,3], and Edward R. Vrscay[1](✉)

[1] Department of Applied Mathematics, Faculty of Mathematics,
University of Waterloo, Waterloo, ON N2L 3G1, Canada
acheeseman@ece.utoronto.ca, ilona@clientoutlook.com, ervrscay@uwaterloo.ca

[2] The Edward S. Rogers Sr. Department of Electrical & Computer Engineering,
University of Toronto, Toronto, ON M5S 3G4, Canada

[3] Client Outlook Inc., Waterloo, ON N2L 6B5, Canada

Abstract. This paper is concerned with the degradation produced in natural images by JPEG compression. Our study has been basically twofold: (i) To find relationships between the amount of compression-induced degradation in an image and its various statistical properties. The goal is to identify blocks that will exhibit lower/higher rates of degradation as the degree of compression increases. (ii) To compare the above *objective* characterizations with *subjective* assessments of observers.

The conclusions of our study are rather significant in several aspects. First of all, "bad" blocks, i.e., blocks exhibiting greater degrees of degradation visually, have among the lowest RMSEs of all blocks and among the medium-to-highest structural similarity (SSIM)-based errors. Secondly, the standard deviations of "bad" blocks are among the lowest of all blocks, suggesting a kind of "Weber law for compression," a consequence of contrast masking. Thirdly, "bad" blocks have medium-to-high high-frequency (HF) fractions as opposed to HF content.

1 Introduction

The study reported in this paper arose from a collaborative research program involving radiologists as well as a leading international developer of medical imaging software (AGFA Healthcare) [4]. Our goal has been to develop objective – as opposed to subjective – methods of assessing the degree to which medical images from various modalities and anatomical regions can be compressed before their diagnostic quality is compromised. There are two major motivations for this research: (1) To date, recommended compression ratios have been based on experiments in which radiologists subjectively assess the diagnostic quality of compressed images. Subjective experiments are labor-intensive and time-consuming (and therefore expensive). (2) Diagnostic quality is clearly related to visual quality. To date, however, radiologists have had to rely mostly on mean

© Springer International Publishing Switzerland 2016
A. Campilho and F. Karray (Eds.): ICIAR 2016, LNCS 9730, pp. 163–172, 2016.
DOI: 10.1007/978-3-319-41501-7_19

squared error (MSE) and its relative, PSNR, because of their prevalent use in the research literature. It is well known, however, that these measures provide poor assessments of visual quality. For this reason, it is necessary to examine whether more recent image fidelity measures, such as the structural similarity index (SSIM) [2], which are known to provide better assessments of visual quality, could be used in the assessment of diagnostic quality.

In [4], we examined the assessments of a number of image quality measures including SSIM and MSE/PSNR and how well they compared with subjective assessments of radiologists based on data collected in two experiments. Very briefly, SSIM provided the closest match to the radiologists' assessments whereas MSE and PSNR were observed to perform inconsistently.

Here it is important to mention that the above results were obtained from **global** analyses of the images, i.e., subjective assessments and objective measures of **entire images**. Generally, however, a radiologist will often judge a compressed image to be diagnostically unacceptable because of perceived degradations in certain regions or features. For this reason, we also pursued the much more ambitious problem of trying to predict which **local regions/features** of a medical image would demonstrate greater degrees of degradation, possibly the first to lose their diagnostic quality as the compression rate is increased.

Unfortunately, this aspect of the study was not conclusive. One problem was that much of our study at that time focussed on CT brain images which exhibit a rather low degree of variability in terms of structure, at least in the cortical region. Other regions of the body, e.g., the abdomen, which exhibit greater variability, did show some trends. This has led us to an examination of "natural images." e.g., the various (nonmedical) test images employed in the standard image processing literature.

As an illustration, in Fig. 1 below are plotted the degradations produced by JPEG compression of a subset (256) of all (4096) nonoverlapping 8×8 blocks of the standard 512×512 pixel, 8 bpp *Lena* image over the range of quality factors $100 \geq Q \geq 10$. Two different measures of degradation are shown in these plots: (a) MSE and (b) "DSSIM," a distance based on the SSIM measure, defined in Eq. (5) of Sect. 2. As expected, the degradation of blocks with respect to both measures generally increases as Q decreases. However, it is also quite clear that there is a great variation in the rates of degradation. Some blocks, which we shall refer to as "bad blocks", exhibit much higher rates of degradation than others, which we shall refer to as "good blocks."

There is an additional complication, however, in that blocks that are "bad" with respect to one measure, say, RMSE, are not necessarily "bad" with respect to the other. On the left of Fig. 2 are shown plots of the ordered measure pairs $(RMSE(Q), DSSIM(Q))$ for the selected 256 blocks over the range $100 \geq Q \geq 10$. On the right of this Figure is shown a plot of $RMSE$ vs $DSSIM$ errors for all 4096 8×8-pixel blocks of the $Q = 50$ JPEG-compressed *Lena* image. Both plots show that there is rather poor correlation between RMSE and DSSIM error measures. This, however, can be viewed quite positively: If MSE fails to detect "bad" blocks by characterizing them as "good", then DSSIM may

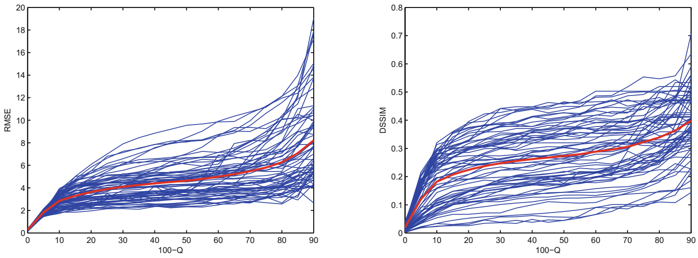

Fig. 1. Degradation vs. $Q' = 100 - Q$ for 256 8×8-pixel blocks of *Lena* image. **Left:** RMSE. **Right:** DSSIM. In both cases, the mean values are also plotted (in red). (Color figure online)

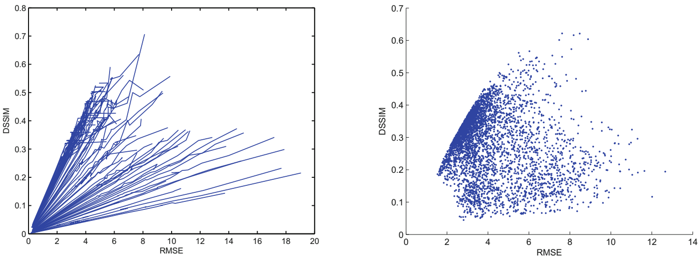

Fig. 2. Left: $RMSE(Q)$ vs. $DSSIM(Q)$ over the range $100 \geq Q \geq 10$ for the 256 8×8-pixel blocks of the *Lena* image. **Right:** $RMSE$ vs. $DSSIM$ errors for all 4096 blocks at $Q = 50$.

characterize them as "bad." In the end, these results must be compared with subjective assessments in order to verify that what is "bad" objectively is also "bad" visually. This is the subject of this preliminary study.

Some obvious questions arise from the above observations, e.g.,

1. What, if any, characteristics of blocks can be used to separate "bad" blocks from "good" blocks **for a given fidelity measure**? Previously we examined standard deviation, total variation, low- and high-frequency content.
2. What, if any, features can be used to characterize blocks that are "bad" with respect to one measure and "good" with respect to the other?
3. Which fidelity measure is **better visually**, i.e., which measure correponds better to human visual perception of degradation?

In our previous studies, none of the characteristics mentioned in Question 1 worked well. In this paper, we show that better indicators are (i) energy and (ii) high frequency **fraction** as opposed to **content**. We have also found that these indicators work equally well for JPEG2000 compression but this will have to be reported elsewhere.

2 Definitions of Important Quantities Used in This Paper

Here we let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{N \times N}$ denote two $N \times N$-dimensional image blocks, i.e., $\mathbf{x} = \{x_{ij}\}$ $1 \leq i, j \leq N$. In this study, $\mathbf{x}$ will usually represent a block of an uncompressed image and $\mathbf{y}$ the corresponding block of the compressed image. The mean squared error/distance (MSE) between $\mathbf{x}$ and $\mathbf{y}$ is given by

$$MSE(\mathbf{x}, \mathbf{y}) = \frac{1}{N^2} \sum_{i,j=1}^{N} (x_{ij} - y_{ij})^2 = \frac{1}{N^2} \|\mathbf{x} - \mathbf{y}\|_2^2 \,, \tag{1}$$

where $\| \cdot \|_2$ denotes the usual Euclidean norm for $N \times N$ matrices. The root mean squared error/distance is

$$RMSE(\mathbf{x}, \mathbf{y}) = \sqrt{MSE(\mathbf{x}, \mathbf{y})} = \frac{1}{N} \|\mathbf{x} - \mathbf{y}\|_2 \,. \tag{2}$$

Of course, $MSE(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.

In this paper, the following form of the structural similarity index (SSIM) [2] between $\mathbf{x}$ and $\mathbf{y}$ is employed,

$$SSIM(\mathbf{x}, \mathbf{y}) = S_1(\mathbf{x}, \mathbf{y}) S_2(\mathbf{x}, \mathbf{y}) = \left[\frac{2\bar{\mathbf{x}}\bar{\mathbf{y}} + \epsilon_1}{\bar{\mathbf{x}}^2 + \bar{\mathbf{y}}^2 + \epsilon_1} \right] \left[\frac{2 s_{\mathbf{xy}} + \epsilon_2}{s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 + \epsilon_2} \right] , \tag{3}$$

where

$$\bar{\mathbf{x}} = \frac{1}{N^2} \sum_{i,j=1}^{N} x_{ij} \,, \; s_{\mathbf{xy}} = \frac{1}{N^2 - 1} \sum_{i,j=1}^{N} (x_{ij} - \bar{\mathbf{x}})(y_{ij} - \bar{\mathbf{y}}) \,, \; s_{\mathbf{x}}^2 = s_{\mathbf{xx}} \,. \tag{4}$$

The small positive constants $\epsilon_1, \epsilon_2 \ll 1$ are added for numerical stability and can be adjusted to accommodate the perception of the human visual system (HVS).

Note that $-1 \leq SSIM(\mathbf{x}, \mathbf{y}) \leq 1$ and $SSIM(\mathbf{x}, \mathbf{y}) = 1$ if and only if $\mathbf{x} = \mathbf{y}$. $SSIM(\mathbf{x}, \mathbf{y})$ is a measure of the similarity between $\mathbf{x}$ and $\mathbf{y}$. In order to be able to make comparisons with the error measures, MSE and RMSE, it is convenient to define a SSIM-based error, or **dissimilarity measure**, as follows,

$$DSSIM(\mathbf{x}, \mathbf{y}) = \sqrt{1 - SSIM(\mathbf{x}, \mathbf{y})} \,. \tag{5}$$

Then $DSSIM(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.

In the case that $\bar{\mathbf{x}} = \bar{\mathbf{y}}$, $S_1(\mathbf{x}, \mathbf{y}) = 1$. This is the case, or very nearly so, when $\mathbf{y}$ is a compressed version of $\mathbf{x}$. The DSSIM distance then becomes [1]

$$DSSIM(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{N^2 - 1}} \frac{\|\mathbf{x}_0 - \mathbf{y}_0\|_2}{\sqrt{s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 + \epsilon_2}} \,, \tag{6}$$

where $\mathbf{x}_0$ and $\mathbf{y}_0$ denote the zero-mean blocks,

$$\mathbf{x}_0 = \mathbf{x} - \bar{\mathbf{x}}\mathbf{1} \quad \mathbf{y}_0 = \mathbf{y} - \bar{\mathbf{x}}\mathbf{1} \,. \tag{7}$$

The $DSSIM(\mathbf{x}, \mathbf{y})$ distance in (6) is generated by a **weighted norm**. DSSIM is seen to penalize blocks $\mathbf{x}$ with lower variance.

Discrete Cosine Transform and JPEG Compression. We let c_{kl}, $0 \leq i, j \leq N-1$, denote the coefficients of the standard DCT of $\mathbf{x} \in \mathbb{R}^{NXN}$ [5]. Since this study is centered around JPEG compression, we consider the special case $N = 8$, where the DCT coeffients are conveniently arranged as an 8×8 array,

$$\mathbf{c} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{07} \\ c_{10} & c_{11} & \cdots & c_{17} \\ \vdots & \vdots & \ddots & \vdots \\ c_{70} & c_{71} & \cdots & c_{77} \end{pmatrix} . \tag{8}$$

From Parseval's Theorem, $\|\mathbf{c}\|_2 = \|\mathbf{x}\|_2$. Now define the following counterdiagonal vectors of the DCT coefficients,

$$\mathbf{d}_m = \{c_{kl},\, k+l = m\}, \quad 0 \leq m \leq 14\,. \tag{9}$$

We define the **low-** and **high-frequency content** of block $\mathbf{x}$ to be as follows,

$$\|\mathbf{x}\|_{lc} = \left[\sum_{m=1}^{6} \|\mathbf{d}_m\|_2^2\right]^{1/2}, \quad \|\mathbf{x}\|_{hc} = \left[\sum_{m=7}^{14} \|\mathbf{d}_m\|_2^2\right]^{1/2}. \tag{10}$$

Note that the DC coefficient $c_{00} = N\bar{\mathbf{x}}$ is omitted from the low-frequency content term since it is generally much greater in magnitude than the other DCT coefficients thereby masking their contributions. Moreover, c_{00} is virtually unchanged by compression. Also note that

$$\|\mathbf{x}\|_{lc}^2 + \|\mathbf{x}\|_{hc}^2 = \|\mathbf{x}\|_2^2 - c_{00}^2 = \|\mathbf{x}_0\|_2^2 = (N^2-1)s_{\mathbf{x}}^2\,. \tag{11}$$

We consider $\|\mathbf{x}_0\|_2$ to define the (reduced) energy of $\mathbf{x}$ and note that it is proportional to the standard deviation of $\mathbf{x}$. We also define the **low-** and **high-frequency fractions** of an image block $\mathbf{x}$ as follows,

$$\|\mathbf{x}\|_{lf} = \frac{\|\mathbf{x}\|_{lc}}{\|\mathbf{x}_0\|_2} \quad \|\mathbf{x}\|_{hf} = \frac{\|\mathbf{x}\|_{hc}}{\|\mathbf{x}_0\|_2}\,. \tag{12}$$

Note that by definition,

$$\|\mathbf{x}\|_{lf}^2 + \|\mathbf{x}\|_{hf}^2 = 1\,. \tag{13}$$

Equation (12) provides a more block-independent, hence compact, characterization of low- and high-frequency content than Eq. (10). A much better idea of the low-high frequency constitution of a block may be obtained by looking at the distribution of low-high fractions of blocks over the quarter circle $x^2 + y^2 = 1$, $0 \leq \theta \leq \frac{\pi}{2}$, as oppposed to low-high content over the first quadrant in $\mathbb{R}^2$.

Finally, because of space limitations, we omit a discussion of JPEG compression since it is a well-known procedure in image processing. The important idea of quantization of the DCT coefficients as determined by the quality factor

Q is discussed in many books, including [5]. Here we simply recall that JPEG compression exploits the fact that the magnitudes of coefficients in the counter-diagonals $\mathbf{d}_m$ generally decrease with m, being very small in the high-frequency region, i.e., $m \geq 7$. It essentially diminishes and, in many cases, removes, high-frequency DCT coefficients of low-magnitude.

3 Quantitative Measure of Compression-Induced Degradation of Image Blocks

We are primarily concerned with the RMSE and DSSIM distances between uncompressed and compressed (8 × 8-pixel) blocks and how they relate to various characteristics of the blocks, including standard deviation, total variation, frequency content and energy. For compactness of presentation, the presentation below is limited to the case of the *Lena* test image. The figures shown below are qualitatively quite similar to those obtained for many other standard "natural" test images, e.g., *Boat*, *San Francisco*, *Peppers*.

Because of space limitations, the figures below show degradation characteristics of subblocks of the *Lena* image compressed with JPEG at quality factor $Q = 50$. This represents a rather mid-range compression level which reveals general characteristics that are seen at both higher and lower compression rates.

In Fig. 3 are presented plots of RMSE and DSSIM errors between uncompressed and JPEG-compressed ($Q = 50$) blocks of the *Lena* test image vs. total variation (TV) of the uncompressed blocks. The left plot demonstrates a quite good correlation between RMSE and TV: blocks with low TV are "good" and those with high TV are "bad". Such a strong correlation is not observed for the DSSIM errors.

In Fig. 4 are shown plots of RMSE and DSSIM compression errors vs. the reduced energies/standard deviations $\|\mathbf{x}_0\|_2$ of the blocks. On the left, we see that blocks with the lowest energy exhibit lowest degradation in terms of RMSE, which is to be expected. The L^2 norms of the high-frequency bands $\mathbf{d}_m$ of these

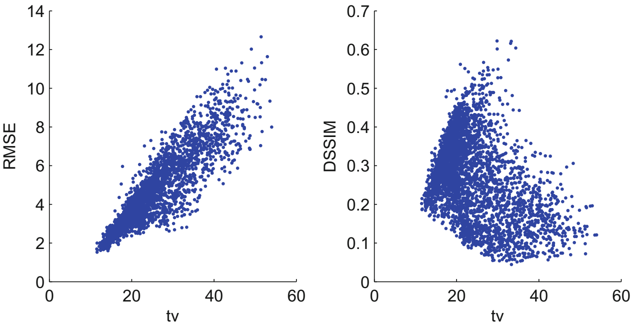

Fig. 3. RMSE and DSSIM distances between 4096 JPEG-compressed ($Q = 50$) and uncompressed 8 × 8-pixel blocks of *Lena* image vs. total variation of the blocks.

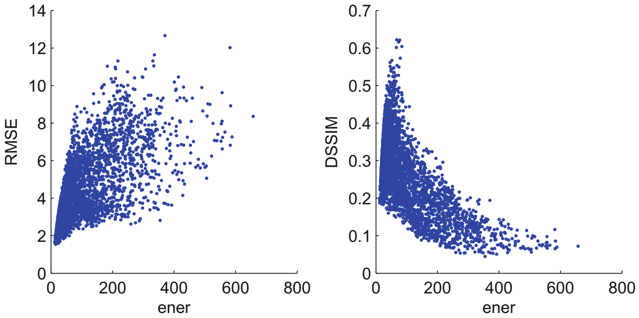

Fig. 4. RMSE and DSSIM distances between 4096 JPEG-compressed ($Q = 50$) and uncompressed 8×8-pixel blocks of *Lena* image vs. energy of the blocks.

blocks will be very small – as such, their removal by JPEG quantization will be virtually negligible. On the other hand, the DSSIM distances exhibit a roughly opposite behaviour – blocks with low energy exhibit a wide range of DSSIM errors, whereas blocks with high energy exhibit low DSSIM errors. This can be explained to a large extent by Eq. (6). These two plots provide a small possibility for separation of RMSE and DSSIM assessments of degradation.

In Fig. 5 are shown plots of RMSE and DSSIM compression errors vs. high-frequency content of the blocks, $\|\mathbf{x}\|_{hc}$ in Eq. (10). Plots of these errors vs low-frequency content, $\|\mathbf{x}\|_{lc}$ in Eq. (10), are virtually identical to the plots of errors vs. energy in Fig. 4 above since most of the energies of the blocks is contained in the low-frequency DCT coefficients.

The three sets of plots presented above show that there is a general correlation between RMSE and the characteristics of total variation (TV), energy (E) and high-frequency content (HC), with the last two being rather clear. Unfortunately, low RMSE does not necessarily imply that the degradations will not be noticed visually. The larger spread of DSSIM errors in the low TV, E and HC regimes

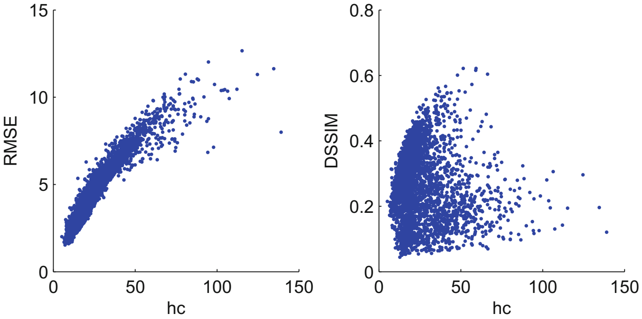

Fig. 5. RMSE and DSSIM distances between 4096 compressed and uncompressed 8×8-pixel blocks of *Lena* image vs. high-frequency content of the blocks.

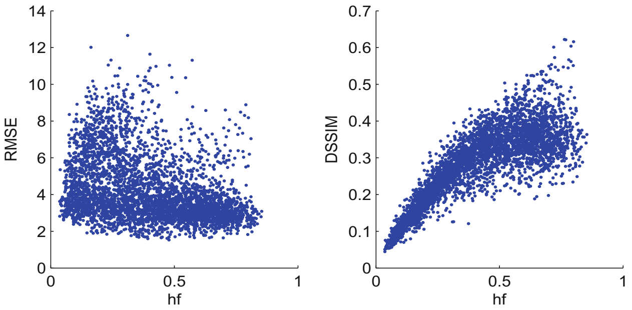

Fig. 6. RMSE and DSSIM distances between 4096 compressed and uncompressed 8 × 8-pixel blocks of *Lena* image vs. high-frequency fractions of the blocks.

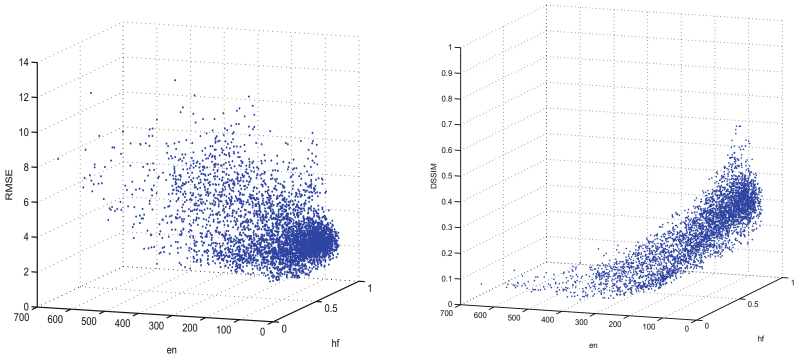

Fig. 7. RMSE and DSSIM distances between 4096 compressed and uncompressed 8 × 8-pixel blocks of *Lena* image vs. high-frequency fractions of the blocks.

indicates that not all blocks with low RMSE are necessarily visually equivalent, i.e., "good" in the sense of DSSIM.

Since JPEG compression generally removes higher-frequency DCT coefficients, it is natural to ask whether blocks with **higher high-frequency fractions**, as opposed to **content**, could exhibit greater **visual degradation**. In Fig. 6 are plotted RMSE and DSSIM compression errors vs. high-frequency fraction, Eq. (12). The plot at the right of this figure is very promising. It shows that that blocks with higher high-frequency fraction exhibit high DSSIM error but, for the most part, low RMSE.

Recall, from Fig. 4, the miniscule separability afforded by the energies of blocks. Figure 7 shows 3D plots of RMSE and DSSIM compression errors vs. both energy and high-frequency fraction. These 3D plots achieve a little more separability of low RMSE/high DSSIM blocks. Figures 4 and 6 represent projections of this 3D plot along in the "hf" and "en" directions, respectively.

4 Comparison with Subjective Evaluations

In order to determine whether the above exercise in separating RMSE and DSSIM assessments is valid visually, we have conducted a set of preliminary subjective experiments involving four individuals. Two images were used in this study, including the 512×512-pixel, 8 bpp *Lena* and *Peppers* images. Each image was JPEG-compressed at four different quality factors, $Q = 10, 15, 25$ and 35. This set of eight compressed images was presented to the subjects in random order several times. The subjects were asked to identify regions of the image that they assessed to be the most noticeably degraded, using an image viewer developed by Mr. Faerlin Pulido, at that time an undergraduate UW Computer Science student. The image viewer allowed the subject to toggle between the compressed image and the uncompressed original image. The subject was able to highlight rectangular regions of the image that he/she assessed as degraded. The coordinates of the blocks in these regions were then imported into MATLAB code for analysis. Our goal was to see how these subjectively-assessed "bad" blocks compared with blocks identified as "bad" by either DSSIM or RMSE. The results obtained by one subject for the *Lena* image compressed at $Q = 25$ are shown in Fig. 8. The results obtained from the other subjects are very similar to these results. Most noteworthy:

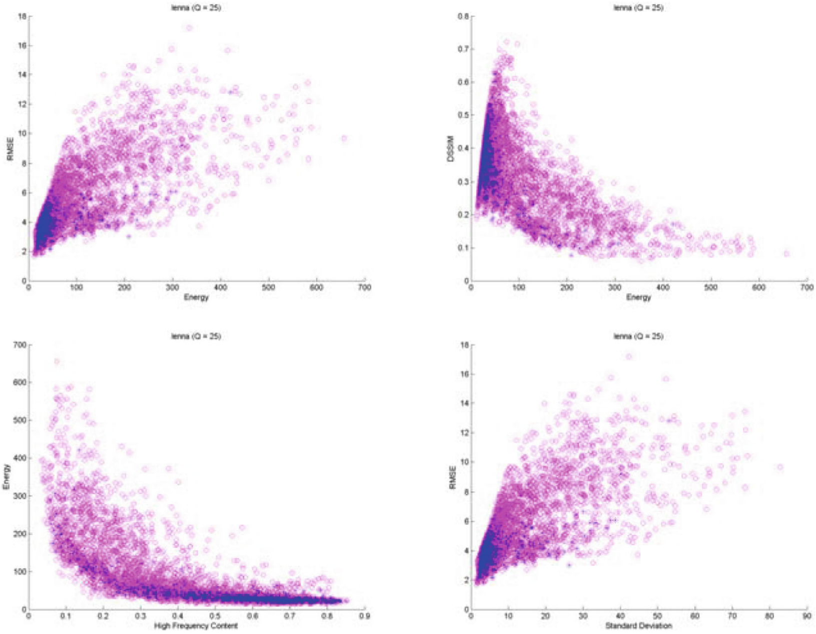

Fig. 8. Results of subjective analysis of JPEG-compressed *Lena* image at $Q = 25$. Red circles denote all 4096 8×8-pixel blocks of image. Blue dots indicate blocks identified by subject as degraded. (Color figure online)

1. The “bad” blocks identified by subjects as visually degraded had **low RMSE** errors and **medium-to-high DSSIM** errors.
2. The “bad blocks” corresponded to uncompressed blocks with **low energies** and **medium-to-high frequency fractions**.

Some interesting conclusions may be made from these features:

1. The fact that blocks with **low RMSE compression error were identified as “bad”** clearly implies that RMSE is **not** a good indicator of degradation.
2. The fact that “bad” blocks correspond to uncompressed blocks with low reduced energy/standard deviation, $\|\mathbf{x}_0\|_2$, indicates that a kind of **perceptual Weber law for compression** is at work here: For a given rate of compression distortions are more likely to be observed for blocks of lower variance. This is actually the principle of “contrast masking,” see, e.g. [3].
3. The fact that “bad” blocks correspond to uncompressed blocks with medium-to-high frequency fraction is quite encouraging. It forces us to break away from the traditional RMSE-centered view that high frequency content is a sufficient criterion for the measurement of degradation.
4. The fact that “bad” blocks are characterized by medium-to-high DSSIM errors, as opposed to low RMSE errors, serves as strong evidence for the need for alternate image quality measures if visual quality is important.

Finally we mention that Comment No. 2 leads to the idea of a variance-based adaptive JPEG compression method which we have developed and which will be reported elsewhere.

Acknowledgements. This research was supported in part by a Discovery Grant (ERV) from the Natural Sciences and Engineering Research Council of Canada.

References

1. Brunet, D., Vrscay, E.R., Wang, Z.: On the mathematical properties of the structural similarity index. IEEE Trans. Image Process. **21**(4), 1488–1499 (2012)
2. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? a new look at signal fidelity measures. IEEE Sig. Process. Mag. **26**(1), 98–117 (2009)
3. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(1), 600–612 (2004)
4. Kowalik-Urbaniak, I.A., et al.: The quest for “diagnostically lossless” medical image compression: a comparative study of objective quality metrics for compressedmedical images. In: SPIE Medical Imaging 2014. doi:10.1117/12:2043196
5. Rao, K.R., Hwang, J.J.: Techniques and Standards for Image, Video and Audio Coding. Prentice Hall, New York (1996)