NISTIR 7939

Effects of Decomposition Levels and Quality Layers with JPEG 2000 Compression of 1000 ppi Fingerprint Images

John M. Libert
Shahram Orandi
Michael D. Garris
John D. Grantham

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 7939

Effects of Decomposition Levels and Quality Layers with JPEG 2000 Compression of 1000 ppi Fingerprint Images

John M. Libert
Shahram Orandi
Michael D. Garris
Information Access Division
Information Technology Laboratory

John D. Grantham
Systems Plus, Inc.
Rockville, MD

August 2013

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary of Commerce for Standards and Technology and Director

Table of Contents

1. INTRODUCTION 1

2. INVESTIGATIVE GOALS 2

3. METHOD 2

3.1. IMAGE DATA 2
3.2. IMAGE COMPRESSION 3
3.2.1. JPEG 2000 Image Compression – Part One 3
3.2.2. JPEG 2000 Image Compression – Part Two 3
3.3. FIDELITY METRICS 4
3.3.1. Count of Changed Pixels 4
3.3.2. Image Mean Squared Error 4
3.3.3. Image Structural Metrics 5
3.3.4. Latent Print Examiner Inspection 8
3.4. HYPOTHESIS TESTING 11
3.4.1. Friedman's Procedure 11
3.4.2. *Post Hoc* Pairwise Multiple Comparisons Procedure 12
3.4.3. Wilcoxon Signed Rank Test 13
3.5. UNCERTAINTY OF SAMPLE STATISTICS 13

4. RESULTS – PART ONE: DECOMPOSITION LEVELS 14

4.1. EXAMINER RATINGS 14
4.2. COMPUTATIONAL IMAGE FIDELITY MEASURES 14
4.2.1. Counts of Changed Pixels 14
4.2.2. Image Mean Squared Error 16
4.2.3. Structural Analysis Based Measures 17
4.2.4. Conclusions: DWT Decomposition Levels 21

5. RESULTS - PART TWO: QUALITY LAYERS 22

5.1. COMPUTATIONAL IMAGE FIDELITY MEASURES 22
5.1.1. Counts of Changed Pixels 22
5.1.2. Image Mean Squared Error 24
5.1.3. Structural Analysis Based Measures 25
5.2. COMPARISON OF RELATIVE ERROR: DECOMPOSITION LEVELS VS. QUALITY LAYERS 29

6. CONCLUSIONS 30

6.1. WAVELET DECOMPOSITION LEVELS 30
6.1.1. To what extent does the number of decomposition levels affect computational measures of image fidelity with respect to source, non-compressed images? 30
6.1.2. Are potential fidelity differences detected by trained latent fingerprint examiners? 30
6.2. EFFECT OF MULTIPLE QUALITY LAYERS 30
6.2.1. What is the effect on image fidelity of using more than a single quality layer? 30
6.2.2. What is the effect on image fidelity of increasing numbers of quality layers? 30

7. REFERENCES 31

List of Tables

Table 1 - Compression Profile – Part One 3
Table 2 - Compression Profile – Part Two 3
Table 3 - Compression Ratio (Quality Layer) Specifications 4
Table 4 - Observation Codes for Compression Degradation Observation 9
Table 5 - Block Design Example 11
Table 6 - Friedman's Test With *Post Hoc* Analysis For Counts of Changed Pixels , SD27 (alpha = 0.05, n=2484, critical difference = 0.1513) 15
Table 7 - Friedman's Test With *Post Hoc* Analysis For Image MSE, FIXT Sample (alpha = 0.05, n=100, critical difference = 0.7540) 16
Table 8 - Friedman's Test With *Post Hoc* Analysis For Image MSE, SD27 (alpha = 0.05, n=2484, critical difference = 0.1513) 16
Table 9 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE – FIXT (alpha = 0.05, n=100, critical difference = 0.7540) 18
Table 10 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE – SD27, (alpha = 0.05, n=2484, critical difference = 0.1513) 18
Table 11 - Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´ – FIXT Sample, (alpha = 0.05, n=100, critical difference = 0.7540) 20
Table 12 - Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´ – SD27, , (alpha = 0.05, n=2484, critical difference = 0.1513) 20
Table 13 - Friedman's Test with *Post Hoc* Multiple Comparisons – Counts of changed Pixels, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013) 22
Table 14 - Friedman's Test with *Post Hoc* Multiple Comparisons – Counts of changed Pixels, SD27 (alpha = 0.05, n=2484, critical difference = 0.2410) 22
Table 15 - Friedman's Test with *Post Hoc* Multiple Comparisons – Image MSE, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013) 24
Table 16 - Friedman's Test with *Post Hoc* Multiple Comparisons – Image MSE, SD27 (alpha = 0.05, n=2484, n=2484, critical difference = 0.2410) 24
Table 17 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013) 26
Table 18 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE, SD27 (alpha = 0.05, n=2484, n=2484, critical difference = 0.2410) 26
Table 19 Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013) 27
Table 20 - Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´, SD27 (alpha = 0.05, n=2484, n=2484, critical difference = 0.2410) 28
Table 21 - Wilcoxon Tests of Extreme Error of Decomposition Levels vs. that for Quality Layers (alpha = 0.05, n=2484) 29

List of Figures

Figure 1 - NIST SIVV metric applied to non-compressed original image and to its corresponding decoded JPEG 2000 compressed image. 5
Figure 2 - Split-Screen Presentation of Image Pairs 10
Figure 3 - Median number of changed pixel values with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images. (Note Median Counts values on the axis are times 10^4.) 15
Figure 4 - Median of image MSE with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images. (CI = confidence interval) 17
Figure 5 - Median SIVV RMSE with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images. (Note that median values on ordinate are multiplied by 10^{-3}.) 19
Figure 6 - Median MSSIM´ with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images. 21
Figure 7 - Median of counts of changed pixels with 95 % confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images. 23
Figure 8 - Median of image MSE with 95 % confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images. 25
Figure 9 - Median of image SIVV RMSE with 95 % confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images. 27
Figure 10 - Median of MSSIM´ with 95% confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images. 28

EXECUTIVE SUMMARY

The criminal justice communities throughout the world exchange fingerprint imagery data primarily in 8-bit gray-scale and at 500 pixels per inch[1] (ppi) or 19.7 pixels per millimeter (ppmm). The Wavelet Scalar Quantization (WSQ) fingerprint image compression algorithm is currently the standard algorithm for the compression of 500 ppi fingerprint imagery. WSQ is a "lossy" compression algorithm. Lossy compression algorithms employ data encoding methods which discard (lose) some of the information through the encoding process in order to achieve more aggressive reduction in the size of the data being compressed. While previous research [29] has shown trained fingerprint examiners tolerant to even extreme loss of image fidelity in performing identification tasks, it remains desirable to ensure that loss of fidelity is kept to a minimum with respect to current or anticipated extensions to identification applications. The WSQ algorithm allows for users of the algorithm to specify how much compression is to be applied to the fingerprint image at the cost of increasingly greater loss in fingerprint image fidelity as the effective compression ratio is increased. The WSQ Gray-Scale Fingerprint Image Compression Specification [1] provides guidance based on an International Association for Identification (IAI) study [2] to determine the acceptable amount of fidelity loss due to compression in order for a WSQ encoder and decoder to meet FBI certifications. These certifications are designed to ensure adherence to the WSQ specification and thereby to ensure fidelity and admissibility in courts of law for images that have been processed by such encoders and decoders.

NIST is in the process of creating a normative guidance on utilizing JPEG 2000 [4] for the compression of fingerprint imagery at 1000 ppi (39.4 ppmm). The normative guidance will provide a comprehensive compression profile for 1000 ppi fingerprint imagery using JPEG 2000, particularly in its specification of software parameters for control and structure of the JPEG 2000 code stream. This study explores and validates two of the recommended parameters for control of the JPEG 2000 code stream. Specifically, it examines the effects of the number of discrete wavelet transform (DWT) decompositions and the use of multiple quality layers (compression rates) used in the encoding.

While investigating each of the two compression parameters separately in this study, fidelity losses are measured via ratings by certified latent fingerprint examiners and by a suite of computational fidelity metrics. Fidelity losses from either decomposition level or quality layer configuration are found to be statistically undetectable among ratings of trained fingerprint examiners. More sensitive tests, computational fidelity metrics, indicated minimum losses with either five or six Discrete Wavelet Transform (DWT) decomposition levels over the six tested in the range from three to eight levels. Fidelity loss due to encoding at multiple quality layers was statistically significant over that of a single compression rate, but such losses were determined to be below the threshold of visibility for trained fingerprint examiners.

[1]Resolution values for fingerprint imagery are specified in pixels per inch (ppi) throughout this document. This is based on widely used specification guidelines for such imagery and is accepted as common nomenclature within the industry. SI units for these will be presented only once.

Abstract

As part of NIST's research efforts to support development of the FBI Next Generation Identification (NGI) systems, this study evaluates effects on image fidelity of wavelet transform decomposition and quality layer options for JPEG 2000 compression of 1000 ppi fingerprint images. First, a suite of 1000 ppi fingerprints are subjected to encoding specifying from three to eight levels of DWT decomposition. Decoded images are compared to non-compressed source images and fidelity loss is evaluated by certified latent fingerprint examiners as well as by several automated computational fidelity metrics. Fidelity losses among the six decomposition level options result in no statistically significant differences among assessments of image degradation by trained fingerprint examiners. Computational metrics find statistically significant fidelity differences, with lowest error for five and six DWT decompositions. In a second experiment JPEG 2000 encoding using version 1.4 of the openJPEG2000 CODEC is applied using from one to nine quality layers. Computational fidelity metrics find statistically significant increase in error with the addition of even a single additional compression rate beyond using only a single target compression ratio of 10:1. Moreover, fidelity loss increases with the addition of each additional quality layer. Hence, while the provision of JPEG 2000 for multiple compression rates within a single code stream adds to the flexibility of these CODECs, the additional flexibility may not be without measurable loss in fidelity. However, maximum error rates measured among the various quality layer configurations are found to be below that detectable by certified fingerprint examiners. Hence, use of quality layers for JPEG 2000 compression of 1000 ppi fingerprints should be limited in number and chosen for anticipated applications.

1. Introduction

NIST is involved in developing a JPEG 2000 profile and recommendations for the compression of 1000 ppi fingerprint images. In support of the FBI's Next Generation Identification (NGI) program a series of studies have been completed toward answering questions regarding compression rate for the 1000 ppi images and optimum method by which to downsample these images to maintain compatibility with legacy 500 ppi systems.

JPEG 2000 [4][1][5][6][7] CODECs provide for the encoding of an image at multiple spatial resolution levels corresponding to the sub-bands of the discrete wavelet transform (DWT) applied to the image and encoding the image at multiple bit rates, i.e. quality layers, in the same code stream. Both of these provisions provide for scalability of both resolution and image quality within the same compressed code stream.

Details of the manner in which JPEG 2000 provides for scalability are elaborated in [8]. Resolution scalability is provided for by the multi-level dyadic (pyramidal) wavelet transform. The (9,7) floating point wavelet [9] may be used for lossy compression or the (5,3) integer wavelet [10] for lossless compression. In either case, a single code stream compressed using an L-level wavelet transform enables reconstruction of images at L + 1 spatial resolutions, each half the dimensions of the preceding. For example, an image initially sampled at 1000 ppi may be compressed at 10:1 using six decomposition levels specified to the CODEC. Upon decompression the image may be reconstructed at the original spatial resolution of 1000 ppi, but the same code stream can also yield an images at 500 ppi, 250 ppi, 125 ppi, approximately 62 ppi, and approximately 31 ppi.

Quality or bitrate scalability is available as an option at encoding. Nominally the JPEG 2000 encoder generates a code stream enabling reconstruction of the input image at some target bit rate or compression level. However, JPEG 2000 CODECs also provide for optional input of a list of bit rates to enable reconstruction from a single code stream images at any of the bit rates contained in the list provided to the encoder. This dimension of scalability together with resolution scalability enables the same code stream to support multiple applications and platforms of varying capability to extract image information at the most appropriate spatial scale and quality.

2. Investigative Goals

This study attempts to explore and validate two of the recommended parameters for control of the JPEG 2000 code stream. Specifically, it examines the effects of the number of discrete wavelet transform (DWT) decompositions and the use of multiple quality layers (compression rates) used in the encoding with the following specific investigative goals:

1a. Expert Examiner Subjective Assessment of Various Wavelet Decomposition Level Counts: Investigate expert examiner opinion to gauge perceived degradation of 1000 ppi fingerprint images as a result of employing different numbers of wavelet decomposition levels during compression.

1b. Signal Analysis and Assessment of Various Wavelet Decomposition Level Counts: Investigate the effects of employing different numbers of wavelet decomposition levels in compression of 1000 ppi fingerprint images via various signal analysis methods.

2a. Expert Examiner Subjective Assessment of Employing Multiple Quality Layers: Investigate expert examiner opinion to gauge perceived degradation of 1000 ppi fingerprint images as a result of employing different numbers of wavelet decomposition levels during compression.

2b. Signal Analysis and Assessment of Employing Multiple Quality Layers: Investigate the effects of employing multiple quality layers in compression of 1000 ppi fingerprint images via various signal analysis methods.

3. Method

The study is divided into two parts. Part One examines the impact on image fidelity of the selection of various numbers of decomposition levels used in encoding the image. Part Two evaluates the effect on image fidelity of compressing images when specifying more than a single quality layer (compression rates) to be included in the codestream.

3.1. Image Data

Two datasets were used for the investigations. One dataset consisted of fingerprint impressions 1 to 12 of the NIST Special Database SD27 [11]. This subset of the SD27 dataset consisted of 2484 images scanned from standard FBI fingerprint 10-print cards at 1000 ppi. These images consisted of 12 fingerprint impressions for each of 207 subjects. Prints 1 to 10 are rolled impressions of right and left hands. Prints 11 and 12 are flat impressions of the right and left thumbs.

For visual rating by trained fingerprint examiners used in part of the study as will be detailed below, a random sample of 100 impressions was selected from the larger SD27 database including only rolled impressions, i.e. from among finger positions 1 to 10. This dataset is referred to as the "FIXT dataset" or "FIXT sample," where FIXT refers to the **F**ingerprint **I**mage e**X**amination **T**ool used in several other NIST studies [29][30] involving fingerprint comparisons by expert fingerprint examiners.

Compression via version 1.4 of the openJPEG2000 CODEC [12][13] was applied to source images of both the SD27 subset and FIXT sample fingerprint datasets using compression profiles to be detailed in sections 3.2.1 and 3.2.2 for decomposition levels (Part One) and quality layers (Part Two) of the study. Computational fidelity metrics (see sections 3.3.1 - 3.3.3) were applied to all processed images for Parts

One and Two of the study. Part One, decomposition level experiments, included comparison of source to processed images by expert fingerprint examiners as described in section 3.3.4.

3.2. Image Compression

3.2.1. JPEG 2000 Image Compression – Part One

All source images were compressed using version 1.4 of the openJPEG2000 [12] CODEC modified by NIST and included in its NIST Biometric Image Software (NBIS) [13]. The target compression rate for all images was 10:1 the profile shown in Table 1. Each source image is compressed using each of the six decomposition levels shown in Table 1.

Table 1 - Compression Profile – Part One

Parameter	Value
Progression Order	RPCL
Code Block Size	64 x 64
Transformation Filter	9, 7 Irreversible
Subsampling Factors	1, 1 (no subsampling)
Tile Size	Image width, image height (single tile for image)
Quality Layers	{10,15,24,34,58,86,144,214}
Decomposition Levels	Each of the set {3,4,5,6,7,8} DWT levels

3.2.2. JPEG 2000 Image Compression – Part Two

Images were encoded using the openJPEG2000 encoder [12] using the profile summarized in Table 2. Only a single decomposition level was used for this part of the study and target compression ratio was 10:1 as with Part One. However, each source image was compressed using each of the nine rate specifications shown in Table 3.

Table 2 - Compression Profile – Part Two

Parameter	Value
Progression Order	RPCL
Code Block Size	64 x 64
Transformation Filter	9, 7 Irreversible
Subsampling Factors	1, 1 (no subsampling)
Tile Size	Image width, image height (single tile for image)
Quality Layers	1 to 9 layers as shown in Table 3
Decomposition Levels	6

Table 3 - Compression Ratio (Quality Layer) Specifications

Layers	Compression Ratios (X:1)	Approximate Bit Rate
9	10,15,24,34,58,86,144,214,324	0.8,0.53,0.33,0.245,0.14,0.09,0.06,0.04,0.03
8	10,15,24,34,58,86,144,214	0.8,0.53,0.33,0.245,0.14,0.09,0.06,0.04
7	10,15,24,34,58,86,144	0.8,0.53,0.33,0.245,0.14,0.09,0.06
6	10,15,24,34,58,86	0.8,0.53,0.33,0.245,0.14,0.09
5	10,15,24,34,58	0.8,0.53,0.33,0.245,0.14
4	10,15,24,34	0.8,0.53,0.33,0.245,
3	10,15,24	0.8,0.53,0.33
2	10,15	0.8,0.53
1	10	0.8

3.3. Fidelity Metrics

In all cases, image fidelity is measured as a comparison between the non-compressed source image and a decoded compressed code stream. Computational fidelity metrics were applied to the 2,484 images of the SD27 dataset paired with each of six compressed versions encoded using decomposition levels as described above. The metrics also were applied to the subset of the images subjected to inspection by trained fingerprint examiners.

3.3.1. Count of Changed Pixels

This metric was slightly different between Parts One and Two of the study.

Counts of Changed Pixels – Part One

Given a source image **X** and a decoded compressed image $\mathbf{Y}^d$, d = one of the decomposition levels {3, 4, 5, 6, 7, 8} count the number of pixels that differ between the two images by one or more gray levels excluding those that have similarly changed for all six decomposition level specifications.

Counts of Changed Pixels – Part Two

As in Part One of the study we count the number of pixels found to have changed from the original (source image) value by one or more gray levels. In analysis of the effect of quality layers, we do no additional filtering of the total count. That is, we do not exclude pixels observed to have changed under all conditions. Hence, given a source image, **X,** and a reconstructed image $\mathbf{Y}^d$, d = 1…9 designating one of the quality level specifications, count the number of pixels that differ between the two images by one or more gray levels.

3.3.2. Image Mean Squared Error

Given a source image, **X**, and a processed image, Y^d, a commonly used measure of image fidelity, the mean squared error is defined as

$$MSE^d = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbf{X}_{i,j} - \mathbf{Y}_{i,j}^d\right)^2}{NM} \quad (1)$$

where *d* is the nominal designation of one of *k* processes or treatments (*d*= 1… *k*), *i, j* are indices to image pixels, and *N* and *M* are numbers of image rows and columns, respectively. Note that the source image and processed image must be equal in dimensions in order to apply this metric.

3.3.3. Image Structural Metrics

NIST Spectral Image Validation/Verification (SIVV) Measures

Developed initially as a method to screen fingerprint databases for non-fingerprint images, segmentation errors, or mislabeled sample rates, the Spectral Image Validation Verification (SIVV) metric [25] provides a comparatively straightforward method by which to assess the frequency structure of an image. Pairwise display of the SIVV signals of non-compressed and compressed images enables summary visualization of the effects of compression across the composition frequency spectrum of the image. As a 1-dimensional representation of a 2-dimensional Fourier spectrum, the SIVV metric applied to a fingerprint image exhibits a peak corresponding to the frequency of the ridge spacing. Also, as shown in Figure 1, comparison of SIVV signals of non-compressed and compressed images shows the loss or gain of power over various frequencies.

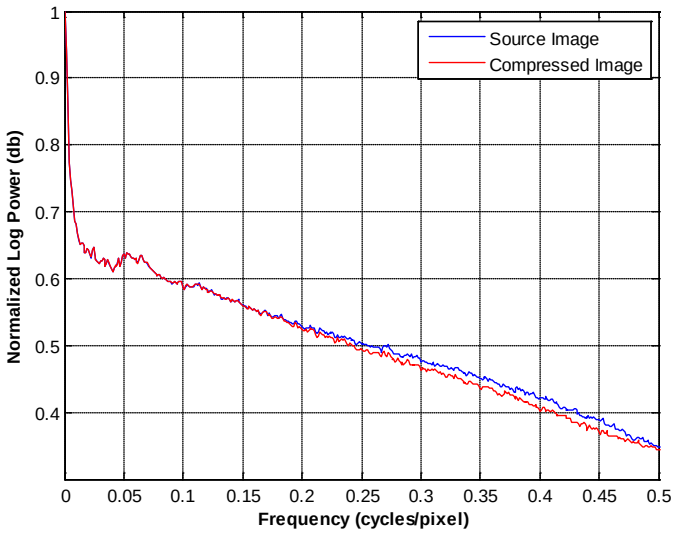


Figure 1 - NIST SIVV metric applied to non-compressed original image and to its corresponding decoded JPEG 2000 compressed image.

The SIVV signals denoted as $\mathbf{s}_1$ and $\mathbf{s}_2$ are respectively vectors of SIVV signal values for images having undergone different processing methods under study. The frequency samples, *f*, in units of cycles per pixel correspond to image pixels or Fourier transform frequencies along the length of one half of the minimum dimension of the 2D Fourier transform of the image under examination. Frequency along this dimension is scaled to the interval [0, 0.5] cycles/pixel. Note that the power value at $f = 0$ is the "direct current" (DC) term, corresponding to the average intensity of the image and is used to normalize the power spectrum.

RMS Error of SIVV Signals

Either differences or ratios of SIVV signals can provide quantitative measures for the comparison of compression methods. For the present study, we examine image differences between pairs of images, $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, with respect to the Root Mean Squared Error (RMSE) between their two SIVV signals, $\mathbf{s}_1$ and $\mathbf{s}_2$, over the entire frequency range 0 - 0.5 cycles/pixel.

$$RMSE(\mathbf{s}_1, \mathbf{s}_2) = \sqrt{\frac{\sum_{i=1}^{n} (\mathbf{s}_{1,i} - \mathbf{s}_{2,i})^2}{n}} \tag{2}$$

where $n = |\mathbf{s}_1| = |\mathbf{s}_2|$ (i.e., the lengths of the signal vectors).

The RMSE metric defined above can provide a measure of the overall difference between the SIVV spectra of the non-compressed source image and that compressed at one of the decomposition levels. In addition to global effects, the RMSE may be evaluated over smaller frequency intervals enabling the comparison of effects over frequency bands that may have particular relevance to fingerprint image quality or matching, as well as quantifying and isolating changes confined to bands that specifically impact either the machine matcher or expert examiners.

Mean Structural Similarity Index (MSSIM)

The Structural Similarity Index (SSIM) , detailed in [22] provides a means by which to compare two images, **X** and **Y**, with respect to structural features of the image without the effects of global changes in luminance or contrast. The SSIM is computed for each of some number of M rectangular subsamples of two images being compared. The SSIM combines measures of local luminance, contrast, and structure.

Given corresponding patches or blocks, **x** and **y**, extracted from two images to be compared, compute the mean of **x** and mean of **y** as

$$\begin{aligned} \mu_{\mathbf{x}} &= \frac{1}{N} \sum_{i=1}^{N} x_i \\ \mu_{\mathbf{y}} &= \frac{1}{N} \sum_{i=1}^{N} y_i \end{aligned} \tag{3}$$

Compute the standard deviation of each of the corresponding blocks, **x** and **y**, as

$$\begin{aligned} \sigma_{\mathbf{x}} &= \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_{\mathbf{x}})^2 \right)^{\frac{1}{2}} \\ \sigma_{\mathbf{y}} &= \left(\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_{\mathbf{y}})^2 \right)^{\frac{1}{2}} \end{aligned} \tag{4}$$

Compute the variance between the corresponding image blocks as

$$\sigma_{\mathbf{xy}} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \mu_{\mathbf{x}}\right)\left(y_i - \mu_{\mathbf{y}}\right) \tag{5}$$

The SSIM for each block of the two images is defined as

$$SSIM(\mathbf{x},\mathbf{y}) = \frac{\left(2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C_1\right)\left(2\sigma_{\mathbf{xy}} + C_2\right)}{\left(\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + C_1\right)\left(\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + C_2\right)}. \tag{6}$$

where C_1 and C_2 are included to handle small values of the image means or standard deviations (which might otherwise result in divide by zero problems) and are computed as

$$\begin{aligned} C_1 &= (K_1 L)^2, \quad K_1 = 0.01 \\ C_2 &= (K_2 L)^2, \quad K_2 = 0.03 \end{aligned} \tag{7}$$

and L corresponds to the dynamic range of pixel values, which for the 8 bit image L=255. The values for K_1 and K_2 may be adjusted, but the values given in (7) were used in the present investigation. Further discussion of the constants, C_1 and C_2 may be found in [22].

Images are processed using overlapping sampling windows. For the present investigation the images were processed using overlapping 11 x 11 pixel blocks, each block centered at a pixel location[2]. It may be appreciated that processing a pair of images in this fashion results in a mapping of the SSIM corresponding roughly to the original image dimensions minus some edge pixels where the sampling window extends beyond the edge of the images. In order to avoid "blocky" artifacts in the SSIM map, the authors of [22] use an 11 x 11 circularly symmetric Gaussian weighting function, $\mathbf{w} = \{w_i \mid i = 1, 2, ..., N\}$, N = 121, with standard deviation of 1.5 pixels and normalized to unit sum. Thus, the estimates of local statistics become

$$\begin{aligned} \mu_{\mathbf{x}} &= \sum_{i=1}^{N} w_i x_i \\ \mu_{\mathbf{y}} &= \sum_{i=1}^{N} w_i y_i \end{aligned} \tag{8}$$

$$\begin{aligned} \sigma_{\mathbf{x}} &= \left(\sum_{i=1}^{N} w_i \left(x_i - \mu_x\right)^2\right)^{\frac{1}{2}} \\ \sigma_{\mathbf{y}} &= \left(\sum_{i=1}^{N} w_i \left(y_i - \mu_y\right)^2\right)^{\frac{1}{2}} \end{aligned} \tag{9}$$

[2] In our application, SSIM outputs are smaller than image inputs, trimmed to dimensions allowing filtering without extending edges of input images.

$$\sigma_{\mathbf{xy}} = \sum_{i=1}^{N} w_i \left(x_i - \mu_x\right)\left(y_i - \mu_y\right) \tag{10}$$

If, as in the present investigation, one requires a single overall measure of fidelity for the entire image pair, the mean SSIM index (MSSIM) is computed as

$$MSSIM\left(\mathbf{X},\mathbf{Y}\right) = \frac{1}{M}\sum_{j=1}^{M} SSIM\left(\mathbf{x}_j,\mathbf{y}_j\right) \tag{11}$$

where M is the number of samples in the SSIM map.

All other computational fidelity metrics detailed in this section indicate fidelity loss increasing from zero. The *MSSIM* shows increasing fidelity loss with decrease of the measure from 1.0. Hence, in order to maintain consistency in the statistical analysis we define a measure MSSIM´ as

$$MSSIM'\left(\mathbf{X},\mathbf{Y}\right) = 1 - MSSIM\left(\mathbf{X},\mathbf{Y}\right) \tag{12}$$

A MATLAB [23] implementation for the MSSIM is given in [24].

3.3.4. Latent Print Examiner Inspection

Direct visual examination and assessment of compression outputs was only performed for Part One of the study. For Part Two we evaluate effects of quality layers indirectly via comparison of maximum fidelity losses observed due to variation in wavelet decomposition levels with those due to inclusion of multiple quality layers.

Examiner Assessment – Part One

Each decoded image, compressed using one of the six decomposition specifications, was paired with its non-compressed source and presented to exactly three examiners. Each examiner was first asked to determine if the image pair being displayed constitutes a matched pair, i.e., if the two fingerprints could have come from the same finger of the same individual. Their responses could be one of three choices:

- The presented image pair is from the same individual ([positive] Identification decision);
- The presented image pair does not appear to be from the same individual (Non-Identification decision);
- Determination of identity cannot be made (Inconclusive).

In contrast to the application of this procedure in [29], the identification exercise was included in the present experiment mainly for control purposes, as all image pairs were either compressed or non-compressed versions of each source image and no non-mates were included in the pairings presented to the examiners.

Subsequent to their identity determination for the pair of images each examiner then evaluated the image pair on fidelity loss of one of the pair relative to the other. To aid in analysis and quantification of fidelity loss, the examiner's evaluation was collected using a Likert-type response scale [26]. The choices that the examiners were allowed to make are provided in Table 4 below. Observation codes are ordered in ascending order indicating a progressively greater amount of degradation from 1 to 4.

Table 4 - Observation Codes for Compression Degradation Observation

Observation Numeric Code	Description
1	No apparent image quality degradation and the quality of Level 2[3] and Level 3 detail in either image should not cause any difficulty in reaching a conclusive decision of identification or exclusion.
2	A noticeable degradation in the quality of Level 2 or Level 3 detail in either image, but not enough to have a negative impact on reaching a conclusive decision of identification or exclusion, though the amount of time to reach a decision may increase.
3	Level 3 detail quality diminished in either image to the extent that a Level 3 identification is questionable or not possible, and/or is significantly more difficult.
4	Level 2 detail quality diminished in either image to the extent that a Level 2 identification becomes questionable or not possible, and/or is significantly more difficult.

Furthermore, the features summarized in observation code 3 are among those typically used for forensic-level decisions while features summarized in observation code 4 are those used primarily by automated-matchers in rendering a match decision.

Examiner responses were recorded by custom test apparatus consisting of a commodity computer and software designed and developed specifically for this and related NIST studies [29][30]. The examiners were not provided any time limits on their responses.

The 600 image pairs (100 images x 6 decomposition level specifications under test) were queued on each examiner's workstation and their presentation order was shuffled randomly on each of the three workstations. Each pair consisted of a non-compressed source image and a version of the identical image decoded from the JPEG 2000 codestream encoded 10:1 using one of the six decomposition level specifications as described in section 3.2.1. Examiners were not informed as to which fingerprint image had been compressed, nor were they told the nature of any process applied to either of the two fingerprints.

The examiners were provided the basic ability to independently reposition, rotate, invert and zoom in and out of each of the two images from the pair being examined. This provided them with the basic tools that they typically employ in their standard operating environments in performing their duties. While they were provided with basic tools, more advanced assistive technologies normally available to some examiners, such as on-screen feature marking or pixel intensity adjustments such as brightness and contrast were not provided to them in the interest of experimental control.

---

[3] The commonly accepted nomenclature defines Level 1 fingerprint details as the overall friction ridge pattern and flow, Level 2 detail as classic Galton features [27] such as minutiae points, and Level 3 as pores, creases, line shapes, incipient ridges and other non Level 1 or 2 features [28].

The image pairs were presented on split-screen to the examiners in randomized order (see Figure 2), and with randomized screen position (left/right split screen placement) to mitigate potential ordering effects or positional bias. Scientists overseeing the tests were blind to the placement order of the images, as well as to the compression level or other processing parameters of image pairs. These factors were tracked by the test apparatus.

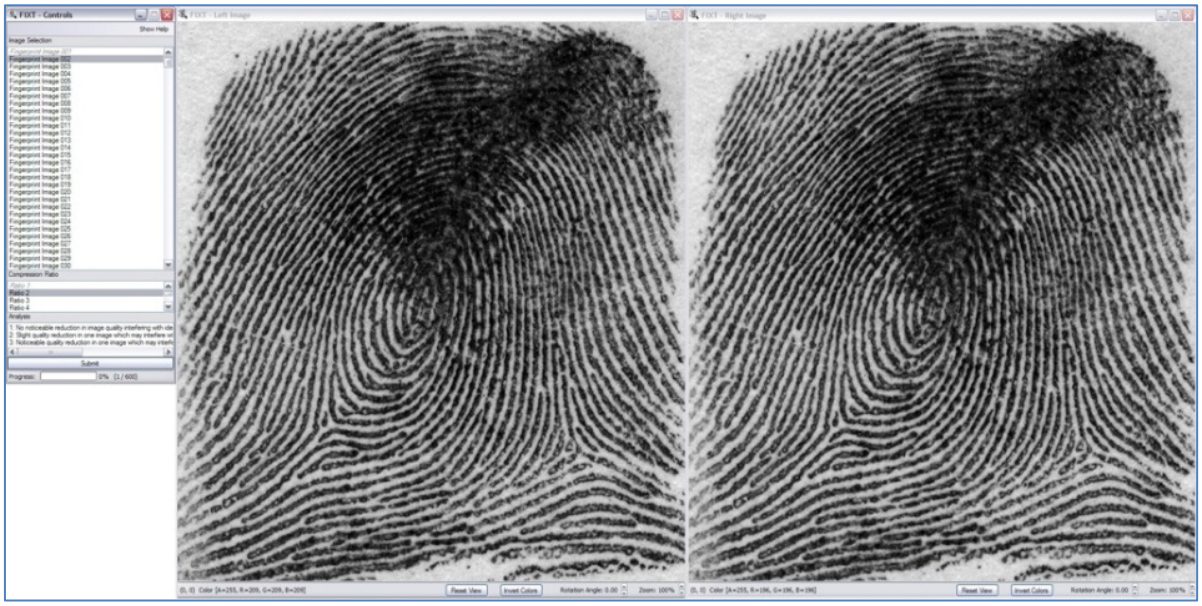


Figure 2 - Split-Screen Presentation of Image Pairs

Each of the 600 image pairs was guaranteed to be observed by 3 different examiners over the course of the study without repeating. Due to the physical limitations of even the best modern monitors, it is impossible to show a 1000 ppi image without zooming or interpolation. The software apparatus enabled the examiner to view images at approximately 10x to 50x of the original size (see [29], Appendix C for additional information).

Once the examiner made a determination for a given pair of images, that pair was marked clearly as complete. The examiner was allowed to return to a completed pair and re-examine that image pair without penalty. The examiner was also allowed to jump to any image pair in the queue regardless of that pair's position in the examiner's queue.

Examiners were provided basic verbal instructions and a demonstration on how to use the workstations, and they were allowed a brief practice session using image pairings that were not part of the study to gain familiarity with the procedure, scoring, and workstation controls. Examiners were allowed to freely ask questions for clarification on the workstations or tasking. The examiners were located in the same room and were allowed to interact freely as they do in their normal professional practice. Finally, the examiners were advised that once they had selected one of the three workstations on which to process images, they

continue to use that workstation exclusively. This was done in order to eliminate the possibility of an examiner processing an image pair more than once.

Examiner Assessment – Part Two

Part Two includes no direct examiner rating measurement. Instead, we use the comparative values of computational fidelity metrics for Part One to infer the likelihood that differences between non-compressed source and encoding using quality layers will be visible to trained fingerprint examiners. If we find, for example, that ratings of trained examiners find no significant difference among the various decomposition level specifications, we can use the fidelity measurements of these images as thresholds below which trained examiners are unlikely to detect differences.

3.4. Hypothesis Testing

3.4.1. Friedman's Procedure

The Friedman test [14][15][16][17] is a non-parametric test for analyzing randomized complete block designs and is an extension of the sign test when there may be more than two treatments of interest. The Friedman test assumes that there are two or more experimental treatments (k). The observations are arranged in b blocks as in the example shown in Table 5.

Table 5 - Block Design Example

Block	Treatment			
	1	**2**	**...**	***k***
1	X_{11}	X_{12}	...	X_{1k}
2	X_{21}	X_{22}	...	X_{2k}
3	X_{31}	X_{32}	...	X_{3k}
...	...	...	...	...
b	X_{b1}	X_{b2}	...	X_{bk}

Let $R(X_{ij})$ be the rank assigned to X_{ij} within block i, and in the case of ties the average rank is used. The ranks are summed to obtain

$$R_j = \sum_{i=1}^{b} R\left(X_{ij}\right) \tag{6}$$

for $j = 1, 2, \ldots, k$, $i = 1, 2, \ldots, b$.

Then the Friedman test is

H_o: Null hypothesis, the treatment effects, τ, have identical effects, i.e. $[\tau_1 = , \ldots, = \tau_k]$

H_a: Alternative hypothesis, at least one treatment is different from at least one other treatment, $[\tau_1, \ldots, \tau_k$ not all equal$]$.

Test Statistic:

$$T_1 = \frac{12}{bk(k+1)} \sum_{i=1}^{k} \left(\left(R_i - b(k+1) \right) / 2 \right)^2 \tag{7}$$

If there are ties, then

$$T_1 = \frac{(k-1)\sum_{i=1}^{k}\left(R_i - \frac{b(k+1)}{2}\right)^2}{A_1 - C_1} \tag{8}$$

where

$$A_1 = \sum_{i=1}^{b}\sum_{j=1}^{k}\left(R\left(X_{ij}\right)\right)^2 \tag{9}$$

$$C_1 = \frac{bk(k+1)^2}{4} \tag{10}$$

Note that Conover [16] recommends the statistic

$$T_2 = \frac{(b-1)T_1}{b(k-1) - T_1} \tag{11}$$

since it has a more accurate approximate distribution. The T_2 statistic is the two-way analysis of variance statistic computed on the ranks $R(X_{ij})$.

The significance level, or the accepted probability of incorrectly rejecting the null hypothesis (i.e., making a Type I error) for the Friedman's test is set at 0.05.

For the T_2 test statistic, the critical region for rejection of the null hypothesis at a chosen type I error rate, α, is

$$T_2 > F_{(\alpha,k-1,(b-1)(k-1))} \tag{12}$$

where $F_{(\alpha,k-1,(b-1)(k-1))}$ is the percent point function of the F distribution.

For the T_1 test statistic, the critical region is given as

$$T_1 = \mathrm{X}^2_{(\alpha,k-1)} \tag{13}$$

where X^2 is the percent point function of the chi-square distribution.

The T_2 approximation is typically preferred as more accurate than T_1.

3.4.2. *Post Hoc* Pairwise Multiple Comparisons Procedure

A statistically significant Friedman's test tells us only that at least one of the treatments (in this case either number of DWT decomposition levels or the number of quality layers used for encoding) differs from at least one other treatment. If the Friedman's test yields a probability greater than or equal to the selected Type I error threshold, the alpha rate, we accept the null hypothesis that there are no differences among the treatments. If the Friedman's test results in a probability value less than the selected alpha rate, we reject the null hypothesis and accept the alternative hypothesis that at least one of the treatments groups has been sampled from a different population. In order to determine which treatments are different, we must perform a *post hoc* analysis of the pairwise differences between the treatments.

We use the Nemenyi test [19][20][32] for *post hoc* multiple pairwise comparisons of treatments with paired observations. This test compares the mean ranks determined by the Friedman's test against critical values of mean rank differences indicative of statistical significance at the selected Type I error rate. Our

use of the Nemenyi test is conservative in that we test all possible comparisons rather than simply each treatment against a control. The Nemenyi test controls the experiment-wise error rate for the multiple comparisons, hence does not require additional adjustment of the alpha level criterion for rejection of the null hypothesis.

We apply the Nemenyi test to the set of mean ranks resulting from the Friedman's test using the following procedure:

1. Order the treatments according to ascending mean rank
2. Given the chosen type I error rate, alpha (α=0.05), number of blocks, *b,* and number of treatments, *k*, decide that treatment, $\tau_u \neq \tau_v$ if

$$\left|R_u - R_v\right| \geq q_\alpha \left[\frac{k(k+1)}{12b}\right]^{1/2} \tag{13}$$

otherwise conclude $\tau_u = \tau_v$, where $u = 1...k$, $v = 1...k$, and $v \neq u$.

Critical values of $q_{\alpha,\nu,k}$ are based on the Studentized range statistic, *Q*. Tabled critical values of this statistic may be found in tables such as that shown in [32] given alpha (α), degrees of freedom (ν), and number of treatments (*k*). Degrees of freedom for large samples (e.g., b > 240) is typically taken as infinite, but may be computed as $\nu = b - k$, where *b* is the number of blocks and *k* is the number of treatments.

Alternately, values of $q_{\alpha,\nu,k}$ may be computed via algorithmic implementations for computation of the cumulative distribution function of the Studentized range statistic and its inverse as was done in the current investigation.

3.4.3. Wilcoxon Signed Rank Test

For one of our analyses we employ the Wilcoxon Signed Rank Test [31][32] [33] to examine the differences between paired observations. This test ranks both positive and negative differences between paired observations. In addition to providing a probability that the two sets of observations cold have come from the same population, the test also reports the number of positive and negative differences that may indicate the direction of differences. Where this test is used in the present study, this information enables us to determine relative effect of one treatment in comparison to another.

3.5. Uncertainty Of Sample Statistics

Where provided, uncertainty of sample statistics, such as the median, is evaluated using a bootstrap procedure. The computation of the sample median is repeated for each of 10,000 replicates of the measurement data drawn randomly with replacement. The standard error, SE, of the statistic is computed as

$$SE = \frac{s_{stat}}{\sqrt{n}} \tag{14}$$

where s_{stat} is the standard deviation of the distribution of values of the replicated statistic, e.g., median of measurements, and *n* is the number of bootstrap replications. For any statistic, the 95 % confidence interval for the estimate is ±1.96(SE). Note that 95 % of the area of the standard normal distribution falls within ±1.96(σ), where σ denotes the standard deviation. Here we model the distribution of the 10,000 medians using the standard normal distribution.

4. Results – Part One: Decomposition Levels

4.1. Examiner Ratings

As may be recalled from section 3.3.4 each image pair was compared by exactly three fingerprint examiners each of whom assigned a score indicating the relative degradation on a four point scale. For analysis, the three scores for each image pair were sorted in ascending order and concatenated to form a composite rating as a three digit integer. A Friedman's test [31] was applied to the 3-digit composite ratings to determine if any of the six decomposition levels showed a statistically significant difference from at least one other decomposition specification. The test yielded a probability estimate of $p = 0.1569$, well above the threshold of $p < 0.05$ required for rejection of the null hypothesis. Accordingly, we conclude that in this test, trained latent fingerprint examiners observed no systematic greater or lesser degradation among the six decomposition level specifications used in the JPEG 2000 compression. It should be noted that the examiner ratings were predominantly rating "2" indicating that the effect of compression to 10:1 was noticeable, but deemed to have no effect on features used in identification (see **Table 4**).

4.2. Computational image fidelity measures

4.2.1. Counts of Changed Pixels

For the FIXT dataset, Friedman's test yielded a probability estimate of $p = 0.5668$ failing to support rejection of the null hypothesis that any of the decomposition levels differed significantly from any other with respect to the number of pixels changed by compression.

The larger SD27 dataset, however, did find statistically significant differences in counts of changed pixels among the six decomposition level specifications, allowing rejection of the null hypothesis at the $p < 0.0001$ level and acceptance of the alternative hypothesis that at least one of the decomposition level specifications generates images that differ from at least one other decomposition level encoding selection. An ordering of the mean ranks observed for each of the decomposition levels is shown in **Table 6**. The ordering of the mean ranks shows counts of changed pixels to increase with decrease in the number of decomposition levels used in the JPEG 2000 encoding. The *post hoc* analysis of pairwise comparisons of the treatments indicates groupings of treatments within which no statistically significant difference was observed. (Note that such lettered groupings of treatments in Table 6 through Table 20 indicate that differences between mean ranks are greater than or equal to the critical values computed for each *post hoc* multiple comparison procedure as described in section 3.4.2.)

As shown in **Table 6**, comparison of counts of pixels changed in intensity value due to compression of the SD27 images finds no significant difference using DWT decomposition levels 8, 7, and 6 at the Type I error rate (alpha) of 0.05. Decomposition levels 7, 6, and 5 differ from 8 levels, but are statistically indistinguishable from 7 and 6 decomposition levels. Decomposition using 4 and 3 levels differ from each other as well as differ with members of groupings A and B.

Table 6 - Friedman's Test With *Post Hoc* Analysis For Counts of Changed Pixels , SD27 (alpha = 0.05, n=2484, critical difference = 0.1513)

Sample	Sum of ranks	Mean of ranks	Groups			
8 Levels	8156.000	3.283	A			
7 Levels	8219.000	3.309	A	B		
6 Levels	8339.000	3.357	A	B		
5 Levels	8536.500	3.437		B		
4 Levels	8973.500	3.613			C	
3 Levels	9940.000	4.002				D

Figure 3 shows the decrease in the median number of image pixels changed via JPEG 2000 compression

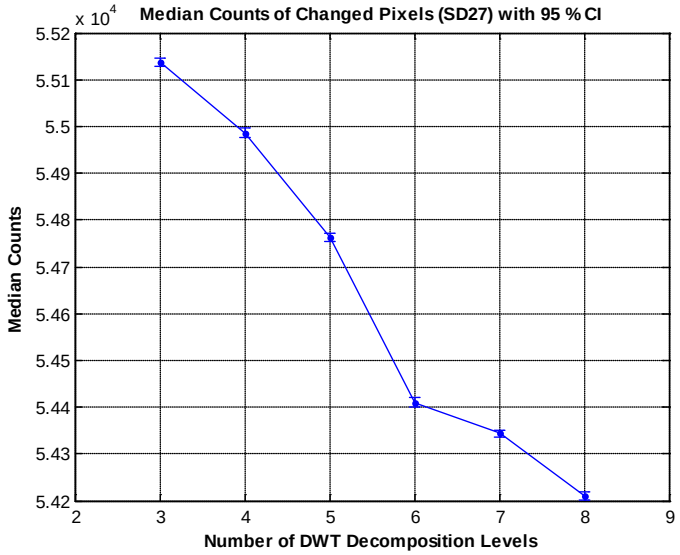


Figure 3 - Median number of changed pixel values with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images. (Note Median Counts values on the axis are times 10^4.)

from three DWT decompositions to eight levels applied to the SD27 1000 ppi source images. The 95 % confidence intervals indicated by the brackets on each of the median values reflect the degree of stability of the median estimate rather than relationship among the decomposition levels. That the confidence intervals of the median estimates do not overlap would suggest significant differences in distributions of measurements for the six decomposition level specifications. However, Friedman's test and the pairwise *post hoc* test provides a better description of the compression behavior on a block-by-block basis. Hence, the *post hoc* multiple comparison tests summarized in **Table 6** indicate that while counts of changed pixels diminishes with increase in decomposition levels, the differences among levels five through seven are not

statistically significant, i.e. mean rank differences among these treatments are less than the critical value for significance at the 0.05 type I error rate.

4.2.2. Image Mean Squared Error

Comparison of image MSE values show the lowest fidelity loss using five decomposition levels, with this specification having the lowest mean rank score for both the 100-sample FIXT image set as well as for the much larger SD27 image collection (see **Table 7** and **Table 8**). No significant difference is found between five and six decomposition levels for the FIXT images, but a significant difference is evident between five and six levels with the SD27 images, with a probability, $p < 0.0001$.

Table 7 - Friedman's Test With *Post Hoc* Analysis For Image MSE, FIXT Sample (alpha = 0.05, n=100, critical difference = 0.7540)

Sample	Sum of ranks	Mean of ranks	Groups
5 Levels	225.000	2.250	A
6 Levels	264.000	2.640	A B
7 Levels	314.000	3.140	B C
4 Levels	329.000	3.290	B C
8 Levels	386.000	3.860	C
3 Levels	582.000	5.820	D

Table 8 - Friedman's Test With *Post Hoc* Analysis For Image MSE, SD27 (alpha = 0.05, n=2484, critical difference = 0.1513)

Sample	Sum of ranks	Mean of ranks	Groups
5 Levels	5205.000	2.095	A
6 Levels	6340.500	2.553	B
7 Levels	8010.500	3.225	C
4 Levels	8166.500	3.288	C
8 Levels	10138.000	4.081	D
3 Levels	14303.500	5.758	E

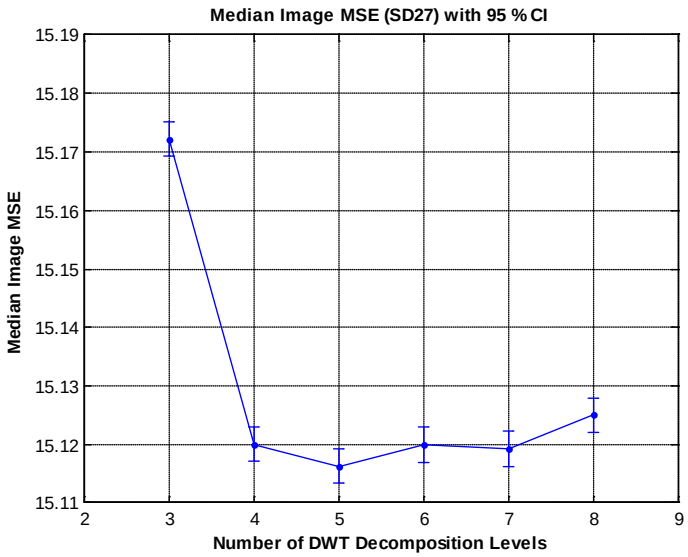


Figure 4 - Median of image MSE with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images. (CI = confidence interval)

In contrast to the situation described above with counts of changed pixels, we see a somewhat different situation with the image MSE measurements. Figure 4 indicates lowest mean squared error with five decomposition levels. This is consistent with the *post hoc* pairwise comparison tests summarized in Table 8 that finds five levels significantly different from all other decomposition level specifications. Overlap among the 95 % confidence intervals on the estimates of the medians indicates that when taken in aggregate, decomposition levels four through seven are closer than indicated by the pairwise comparison tests.

4.2.3. Structural Analysis Based Measures

SIVV RMSE

Comparison of the frequency structures of non-compressed to decoded compressed images show no significant differences using any decomposition level specification higher than three levels, with six levels appearing at the top of the list, i.e. showing lowest fidelity loss, for both the 100-sample FIXT images as well as for the SD27 fingerprints, though **Table 9** and **Table 10** show different ordering of mean ranks for the two datasets.

Table 9 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE – FIXT (alpha = 0.05, n=100, critical difference = 0.7540)

Sample	Sum of ranks	Mean of ranks	Groups		
6 Levels	295.000	2.950	A		
7 Levels	320.000	3.200	A	B	
4 Levels	329.000	3.290	A	B	
5 Levels	353.000	3.530	A	B	
8 Levels	371.000	3.710	A	B	C
3 Levels	432.000	4.320			C

Table 10 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE – SD27, (alpha = 0.05, n=2484, critical difference = 0.1513)

Sample	Sum of ranks	Mean of ranks	Groups	
6 Levels	8509.000	3.426	A	
5 Levels	8532.000	3.435	A	
8 Levels	8541.000	3.438	A	
7 Levels	8551.000	3.442	A	
4 Levels	8818.000	3.550	A	
3 Levels	9213.000	3.709		B

Medians of SIVV RMSE measurements shown in Figure 5 point to five DWT levels as exhibiting the least error in frequency spectrum in comparison to that of source images. Pairwise comparison tests, however, as shown in Table 10 indicate that four through eight levels of decomposition are statistically indistinguishable.

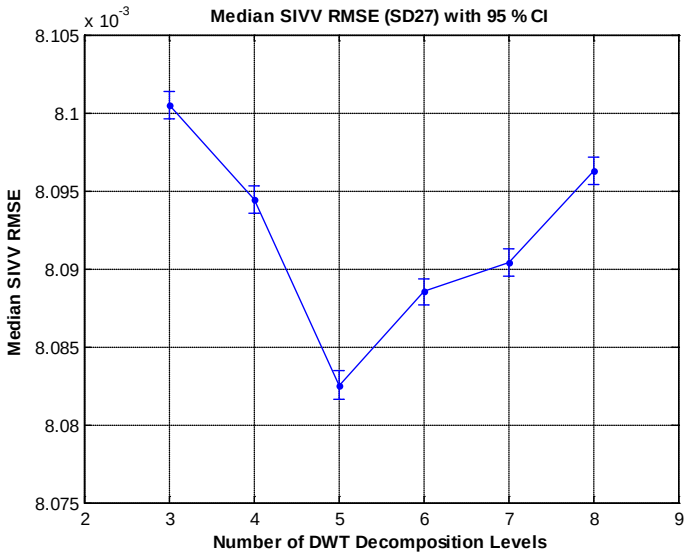


Figure 5 - Median SIVV RMSE with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images. (Note that median values on ordinate are multiplied by 10^{-3}.)

MSSIM´

Our second image structural fidelity metric, MSSIM´, shows five decomposition levels at the top of the list for both the 100-sample FIXT images as well as with the SD27 data[4]. **Table 11** shows a grouping of 5, 6, 7, and 4 levels for the smaller dataset whereas Table 12 again finds five decomposition levels at the top position with a significant difference ($p < 0.0001$) with the next in order, six decomposition levels. This relationship between five and six levels is similar to that observed for the MSE measurements shown in **Table 7** and **Table 8**, possibly due to the inclusion of luminance measures in the MSSIM´ which might be expected to correlate with the MSE measure to an extent greater than that of the SIVV RMSE.

[4] Note that the maximum MSSIM is 1.0, indicating no loss in fidelity. Hence, in contrast to the other fidelity measures used here, mean ranks are ordered highest to lowest for this metric.

Table 11 - Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´ – FIXT Sample, (alpha = 0.05, n=100, critical difference = 0.7540)

Sample	Sum of ranks	Mean of ranks	Groups					
5 Levels	447.000	2.530	A					
6 Levels	421.000	2.790	A					
7 Levels	401.000	2.990	A					
4 Levels	375.000	3.250	A	B				
8 Levels	321.000	3.790		B				
3 Levels	135.000	5.650			C			

Table 12 - Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´ – SD27, , (alpha = 0.05, n=2484, critical difference = 0.1513)

Sample	Sum of ranks	Mean of ranks	Groups					
5 Levels	11695.000	2.292	A					
6 Levels	10821.000	2.644		B				
7 Levels	9556.000	3.153			C			
4 Levels	9173.000	3.307				D		
8 Levels	7796.000	3.862					E	
3 Levels	3123.000	5.743						F

Inasmuch as the MSSIM´ has a maximum value of 1.0, we subtract the values from this maximum to put the medians of this error measure on a scale similar to that for other error metrics. Median and 95 % confidence intervals are plotted in Figure 6. We see that the pattern of medians for this structural error metric is similar to that of the other structural metric, the SIVV RMSE. As with the SIVV RMSE, the median MSSIM´ suggests five decomposition levels as the having the least fidelity loss relative to the source images. In this case, the difference between five levels and the next closest, six decomposition levels, is statistically significant.

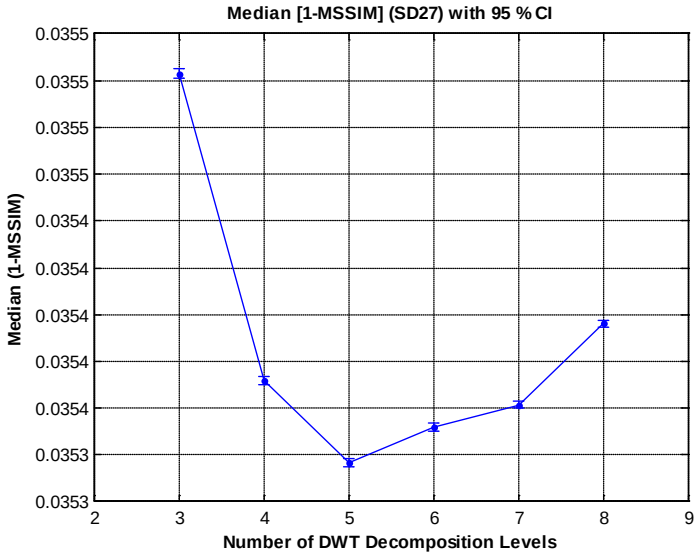


Figure 6 - Median MSSIM´ with 95 % confidence intervals for each of six decomposition level specifications used in JPEG 2000 compression of SD27 1000 ppi source images.

4.2.4. Conclusions: DWT Decomposition Levels

Five and six decompositions levels rank at the top of the list of decomposition level specifications or are found in the top performing group among all the fidelity measurements. As measurements on the SD27 image dataset are more likely to yield stable statistics, five decomposition levels might appear best for JPEG 2000 compression of 1000 ppi images. However, to support additional flexibility in the decoding of compressed codestreams at lower than maximum resolution, six levels could be justified as acceptable as well as five DWT decomposition levels. Lepley [3] suggests that inasmuch as 500 ppi fingerprint images have been compressed via the Wavelet Scalar Quantization (WSQ) CODEC [35] [36] [37] [38] [39] using five decomposition levels, it is reasonable to provide an additional DWT level for images having twice the initial resolution.

5. Results - Part Two: Quality Layers

5.1. Computational Image Fidelity Measures

5.1.1. Counts of Changed Pixels

The Friedman's test comparison of counts of changed pixels showed significant differences among the quality layer specifications for both the 100-sample FIXT images as well as for the SD27 images. *Post hoc* analyses summarized in **Table 13** and **Table 14** show increasing differences from the single quality layer to nine quality layers. For the FIXT sample, the groupings overlap whereas with the larger SD27 dataset, all quality level specifications differ from all others.

Table 13 - Friedman's Test with *Post Hoc* Multiple Comparisons – Counts of changed Pixels, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013)

Sample	Sum of ranks	Mean of ranks	Groups						
1 Layers	191.500	1.915	A						
2 Layers	306.500	3.065	A	B					
3 Layers	418.000	4.180		B	C				
4 Layers	497.000	4.970			C	D			
5 Layers	551.500	5.515				D	E		
6 Layers	587.500	5.875				D	E	F	
7 Layers	609.500	6.095				D	E	F	
8 Layers	656.500	6.565					E	F	
9 Layers	682.000	6.820						F	

Table 14 - Friedman's Test with *Post Hoc* Multiple Comparisons – Counts of changed Pixels, SD27 (alpha = 0.05, n=2484, critical difference = 0.2410)

Sample	Sum of ranks	Mean of ranks	Groups								
1 Layers	3606.000	1.452	A								
2 Layers	6349.000	2.556		B							
3 Layers	9015.000	3.629			C						
4 Layers	11671.500	4.699				D					
5 Layers	13736.500	5.530					E				
6 Layers	15320.500	6.168						F			
7 Layers	16521.500	6.651							G		
8 Layers	17440.500	7.021								H	
9 Layers	18119.500	7.294									I

The Friedman's test considers within block rankings of treatment options, hence such tests may be performed without normalization to adjust for differences in image sizes. Computation of confidence intervals on the median of counts is complicated by the interaction of data resampling with the large variation in image sizes. Hence, for estimation of sample medians, we convert the counts to proportion or percent of pixels by dividing counts by the image size in pixels. Medians of percent pixels changed are thus displayed in Figure 7 with 95 % confidence intervals on the estimates of this statistic. We see the single quality layer as showing the lowest percentage of changed pixels and an almost linear increase with each additional quality layer up to seven layers. beyond seven layers, the rate of increase decreases to almost level off. The overlap of confidence intervals in seven to nine quality layers reflects potential similarity among medians and aggregate similarity of treatments. Such uncertainty among sample statistics does not over-rule comparisons of treatments on a block by block (image by image basis.) Hence, the separation of, and ordering of, the medians remains consistent with the pairwise comparisons summarized in Table 14. We note, however, that while statistically significant, the incremental increase in percent changed pixels is quite small (on the order of only hundredths of a percent) relative to the initial impact of around 44.5 % compressing only to the single 10:1 rate.

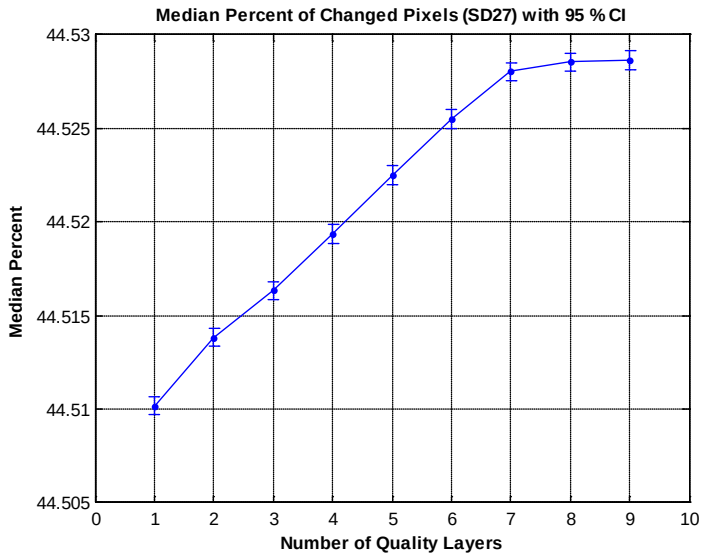


Figure 7 - Median of counts of changed pixels with 95 % confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images.

5.1.2. Image Mean Squared Error

Image MSE measures show fidelity results similar to that of counts of changed pixels with ordering of fidelity loss successively greater with larger numbers of quality layers from one to nine, with some grouping for the smaller FIXT sample (**Table 15**) and complete separation for the SD27 images (**Table 16**).

Table 15 - Friedman's Test with *Post Hoc* Multiple Comparisons – Image MSE, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013)

Sample	Sum of ranks	Mean of ranks	Groups				
1 Layers	123.000	1.230	A				
2 Layers	239.000	2.390	A				
3 Layers	381.000	3.810		B			
4 Layers	471.500	4.715		B	C		
5 Layers	560.000	5.600			C	D	
6 Layers	614.000	6.140				D	E
7 Layers	664.000	6.640				D	E
8 Layers	714.000	7.140					E
9 Layers	733.500	7.335					E

Table 16 - Friedman's Test with *Post Hoc* Multiple Comparisons – Image MSE, SD27 (alpha = 0.05, n=2484, n=2484, critical difference = 0.2410)

Sample	Sum of ranks	Mean of ranks	Groups							
1 Layers	3221.500	1.297	A							
2 Layers	5967.000	2.402		B						
3 Layers	8714.000	3.508			C					
4 Layers	11428.000	4.601				D				
5 Layers	13649.500	5.495					E			
6 Layers	15403.000	6.201						F		
7 Layers	16823.000	6.773							G	
8 Layers	17908.000	7.209								H
9 Layers	18666.000	7.514								I

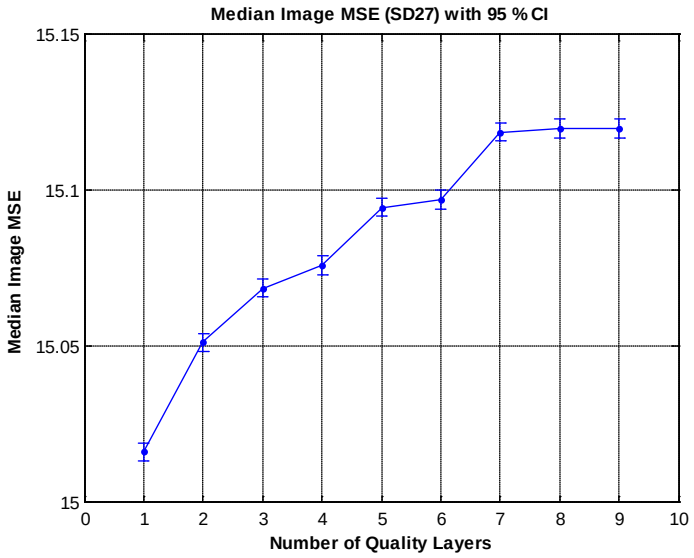


Figure 8 - Median of image MSE with 95 % confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images.

Median MSE of **Figure 8** shows the steepest increase in error in going from the single quality layer to two layers and a leveling off of the increase by seven quality layers. Pairwise comparisons summarized in **Table 16** show all differences among quality layers to be statistically significant, though it appears that when taken as aggregate samples described by median estimates, some of the quality layer specifications appear similar in MSE.

5.1.3. Structural Analysis Based Measures

SIVV RMSE

Differences in the SIVV spectral signals indicate less variability than that seen with the pixel-based measures described above. We find that addition of one quality layer to the single target compression rate to have no significant impact on the power spectrum, but observe additional layers to fall into a cluster of similar degree of degradation.

Table 17 shows the beginning of this pattern with the FIXT sample images, but **Table 18** covering the larger image dataset shows a fully developed pattern of three overlapping clusters in which fidelity decreases with increasing number of quality layers.

Table 17 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013)

Sample	Sum of ranks	Mean of ranks	Groups		
1 Layers	307.000	3.070	A		
2 Layers	374.000	3.740	A	B	
3 Layers	475.000	4.750		B	C
4 Layers	519.500	5.195			C
5 Layers	552.000	5.520			C
6 Layers	562.000	5.620			C
7 Layers	564.000	5.640			C
8 Layers	573.000	5.730			C
9 Layers	573.500	5.735			C

Table 18 - Friedman's Test with *Post Hoc* Multiple Comparisons – SIVV RMSE, SD27 (alpha = 0.05, n=2484, n=2484, critical difference = 0.2410)

Sample	Sum of ranks	Mean of ranks	Groups		
1 Layers	11589.500	4.666	A		
2 Layers	12094.000	4.869	A	B	
3 Layers	12320.000	4.960		B	C
4 Layers	12511.000	5.037		B	C
5 Layers	12538.500	5.048		B	C
6 Layers	12601.000	5.073		B	C
7 Layers	12647.000	5.091		B	C
9 Layers	12739.000	5.128			C
8 Layers	12740.000	5.129			C

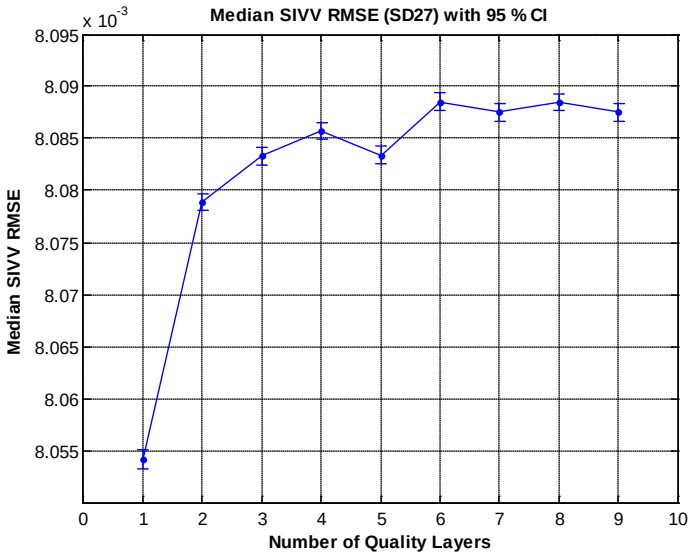


Figure 9 - Median of image SIVV RMSE with 95 % confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images.

The plot of median values (**Figure 9**) of the first of our structural metrics reinforces the contrast between the single quality layer and two or more layers as well as the pairwise similarity among the multiple layer JPEG 2000 specifications.

MSSIM´

With the 100 sample FIXT image set, we find no significant difference in the MSSIM´ between a single quality layer and two layers (see **Table 19**). These two quality layer configurations then differ significantly with several other clusters. No significant differences in MSSIM´ are found from six to nine quality layers.

Table 19 Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´, FIXT Sample (alpha = 0.05, n=100, critical difference = 1.2013)

Sample	Sum of ranks	Mean of ranks	Groups				
1 Layers	877.000	1.230	A				
2 Layers	760.000	2.400	A				
3 Layers	618.000	3.820		B			
4 Layers	530.500	4.695		B	C		
5 Layers	440.000	5.600			C	D	
6 Layers	386.000	6.140				D	E
7 Layers	336.000	6.640				D	E
8 Layers	286.000	7.140					E
9 Layers	266.500	7.335					E

With the much larger SD27 image dataset we find all quality layer specifications significantly different from all others (see **Table 20**).

Table 20 - Friedman's Test with *Post Hoc* Multiple Comparisons – MSSIM´, SD27 (alpha = 0.05, n=2484, n=2484, critical difference = 0.2410)

Sample	Sum of ranks	Mean of ranks	Groups
1 Layers	21604.500	1.303	A
2 Layers	18865.000	2.405	B
3 Layers	16115.000	3.512	C
4 Layers	13407.000	4.603	D
5 Layers	11191.500	5.495	E
6 Layers	9432.000	6.203	F
7 Layers	8032.000	6.767	G
8 Layers	6953.000	7.201	H
9 Layers	6180.000	7.512	I

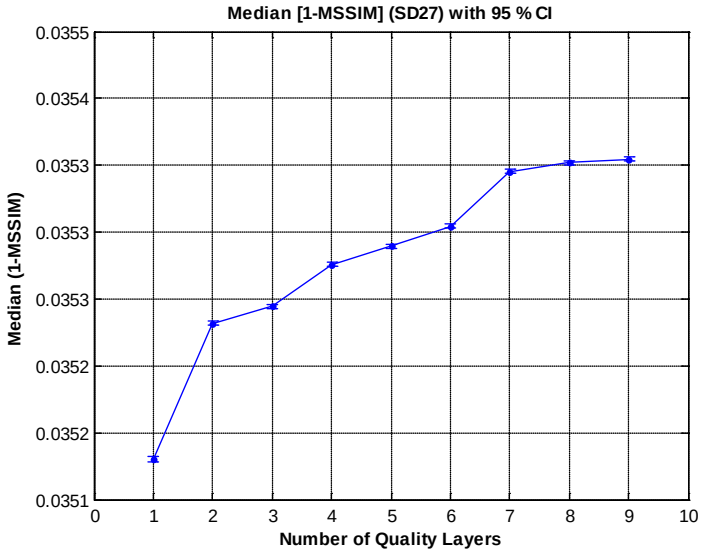


Figure 10 - Median of MSSIM´ with 95% confidence intervals for each of nine quality layer specifications used in JPEG 2000 compression of SD27 1000 ppi source images.

As we mentioned earlier, for purposes of median plots, we subtract the MSSIM´ values from 1.0 such that increasing error is shown as increasing values of the metric. As with the other metrics, we see via **Figure 10**

an increase in error with increasing numbers of quality layers. Pairwise comparisons summarized in **Table 20** show all quality layer configurations to differ significantly from all others. That the median estimates are very stable over the bootstrap replications reinforces the statistical significance observed and that JPEG 2000 compression using quality layers does degrade structural features of the image beyond that observed using only a single target compression rate. In the next section we address whether the magnitude of the observed fidelity degradation should be considered problematic to fingerprint identification via visual examination by trained fingerprint examiners.

5.2. Comparison of relative error: Decomposition Levels vs. Quality Layers

As indicated previously, trained examiners' ratings showed no significant difference among the six DWT decomposition specifications used for encoding fingerprint images. An examiner study was not included in Part Two of the present investigation. Hence, we are unable to directly test the effects of using more than a single quality layer for JPEG 2000 encoding of fingerprints. However, we are able to perform an indirect assessment of the likelihood that image fidelity loss due to use of multiple quality layers might be visible to trained examiners.

We may compare our measures of fidelity loss among the six decomposition levels on an image by image basis to measures similarly computed for the quality layer specifications in order to infer potential fidelity loss as being below, or possibly above, the visual threshold of trained fingerprint examiners.

For this, we consider the 100 sample images from the FIXT dataset used in the examiner portion of the study. For each image, we find the maximum of the values of MSE, SIVV RMSE, and MSSIM´ for the six decomposition levels[5]. We find the corresponding maximum values of these computational metrics for the quality layers, excluding that for the single quality layer.

Table 21 - Wilcoxon Tests of Extreme Error of Decomposition Levels vs. that for Quality Layers (alpha = 0.05, n=2484)

	Differences			
Metric	**Positive**	**Negative**	**Zero**	**Wilcoxon Probability**
MSE	99	0	1	p<0.0001
SIVV RMSE	87	6	7	p<0.0001
MSSIM	84	0	16	p<0.0001

Table 21 summarizes the Wilcoxon Signed Rank Tests of the difference between maximum values of fidelity error metrics for images encoded using each of the six DWT decomposition level specifications and those for corresponding images compressed using quality layer specifications shown in **Table 3**. For each of the metrics we find a significant difference between fidelity of the two JPEG 2000 encoding treatments with $p < 0.0001$ in each case, significantly less than the accepted Type I error rate, $\alpha = 0.05$.

[5] We exclude counts of changed pixels from this analysis as the criteria for counting are different in Part One and Part Two of the study.

As the Wilcoxon Signed Rank Test evaluates the differences between paired observations of the two treatments, the predominantly positive differences of each tests indicate that maximum fidelity loss of decomposition level treatments exceeds that of quality layer treatments with most of the 100 images processed. We are accordingly able to infer that given that trained latent fingerprint examiners could not differentiate among the decomposition levels at measured fidelity error, it is unlikely that images exhibiting lesser error due to encoding with multiple quality layers would be differentiable by the visual examination.

6. Conclusions

6.1. WAVELET DECOMPOSITION LEVELS

6.1.1. To what extent does the number of decomposition levels affect computational measures of image fidelity with respect to source, non-compressed images?

For 1000 ppi fingerprint images, DWT decomposition at five and six levels appears to be a "sweet spot" exhibiting less fidelity loss than fewer, or even greater, decompositions.

6.1.2. Are potential fidelity differences detected by trained latent fingerprint examiners?

While computational fidelity metrics find statistically significant differences among images compressed at different numbers of decomposition levels, the differences appear to be largely insignificant with respect to visual comparison by trained fingerprint examiners.

6.2. EFFECT OF MULTIPLE QUALITY LAYERS

6.2.1. What is the effect on image fidelity of using more than a single quality layer?

Computational measurements show that JPEG 2000 compression using more than a single quality level will result in small, but statistically significant loss in fidelity in comparison of non-compressed source images with those decoded from compressed codestreams. While the fidelity losses are statistically significant, they have no observed effect with respect to use of the once-compressed images for fingerprint identification by certified latent fingerprint examiners.

6.2.2. What is the effect on image fidelity of increasing numbers of quality layers?

Fidelity loss increases with the each additional quality layer. Again, while such losses are statistically significant, they appear to have no observed effect with respect to judgements of certified fingerprint examiners. Moreover, incremental loss in fidelity is small relative to that incurred by compression at a single target rate of, e.g., 10:1. Interestingly, it appears as though the increase in fidelity loss levels out around seven quality layers, with little incremental loss with additional low bit rate layers. Still, however, we do not yet know the effects of even small incremental fidelity loss on automated matching. Hence, it is recommended that selection of quality layers be limited to those having known applications or for which potential applications are anticipated.

7. References

[1] "WSQ Gray-Scale Fingerprint Image Compression Specification" Version 3.1. https://www.fbibiospecs.org/docs/WSQ_Gray-scale_Specification_Version_3_1.pdf. Retrieved 2010-01-11.

[2] Fitzpatrick, M. et al. 1994, "WSQ Compression / Decompression Algorithm Test Report", IAI Annual Conference.

[3] Lepley, M. A. Profile for 1000 ppi Fingerprint Compression. http://www.mitre.org/work/tech_papers/tech_papers_04/lepley_fingerprint/lepley_fingerprint.pdf. Retrieved 2011-01-11.

[4] ISO/IEC 15444-1: JPEG 2000 image coding system, 2000.

[5] ISO/IEC 10918-1, ITU Recommendation T.81: Digital compression and coding of continuous tone still images, requirements and Guidelines, September 1993.

[6] Marcellin, M. W., Gormish, M. J., Bilgin, A., Boliek, M. P. *An Overview of JPEG-2000*, Proc. of IEEE Data Compression Conference, pp. 523-541, 2000.

[7] Taubman, D. S. and Marcellin, M. S. JPEG 2000 Image Compression Fundamentals, Standards, and Practice. Kluwer Academic Publishers Norwell, MA, USA, (2001), 773pp.

[8] Taubman, D., Ordentlich, E., Weinberger, M., Seroussic, G. *Embedded block coding in JPEG 2000*, Signal Processing: Image Communication 17 (2002) 49–72.

[9] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, *Image coding using wavelet transform,* IEEE Trans. on Image Proc., vol. 1, no. 2, pp 205-220, Apr. 1992.

[10] R. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, *Wavelet transforms that map integers to integers*, Journal of Appl. and Comp. Harmon. Analy., vol. 5, pp. 332-369, 1998.

[11] Garris, M. D. and McCabe, R. M. *NIST Special Database 27: Fingerprint Minutiae from Latent and Matching Tenprint Images*, NIST Technical Report NISTIR 6534 & CD-ROM, June 2000.

[12] Open JPEG, Open source software in C for JPEG2000, http://www.openjpeg.org/ .

[13] NIST, NIST Biometric Image Software. http://www.nist.gov/itl/iad/ig/nbis.cfm (Retrieved 05/01/2013).

[14] Friedman, Milton. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". Journal of the American Statistical Association (American Statistical Association) 32 (200): 675–701, (December 1937). doi:10.2307/2279372. JSTOR 2279372.

[15] Friedman, Milton. "A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance". Journal of the American Statistical Association (American Statistical Association) 34 (205): 109, (March 1939). doi:10.2307/2279169. JSTOR 2279169.

[16] Conover, W. J. Practical Nonparametric Statistics, Third Edition, Wiley, 1999, pp. 367-373.

[17] Hollander, M., and Wolfe, D. A. Nonparametric Statistics. (1973). New York: J. Wiley.

[18] Curran-Everett, D. *Multiple comparisons: philosophies and illustrations*. Am J Physiol. Regulatory Integrative Comp Physiol. 279: 8, (2000).

[19] Nemenyi, P.B. (1963) Distribution-free Multiple Comparisons. PhD thesis, Princeton University.

[20] Demšar, J. (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, V. 7, pp. 1 – 30.

[21] Tukey, J. W. Exploratory Data Analysis. Addison-Wesley, Reading. 1977.

[22] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. *Image quality assessment: from error visibility to structural similarity*, IEEE Trans. Image Process. 13(4), 600–612 (2004).

[23] MATLAB Version: 8.1.0.604 (R2013a). Natick, Massachusetts: The MathWorks Inc., 2010.

[24] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. SSIM Index for Image Quality Assessment: MATLAB Source Code, https://ece.uwaterloo.ca/~z70wang/research/ssim/ (Retrieved 05/07/2013).

[25] Libert, J. M., Orandi, S., Grantham, J. A 1D Spectral Image Validation/Verification Metric for Fingerprints (NIST IR 7599), National Institute of Standards and Technology, Gaithersburg, MD, 2009, http://nvlpubs.nist.gov/nistpubs/ir/2009/ir7599.pdf .

[26] Likert, R. (1932). *A Technique for the Measurement of Attitudes,* Archives of Psychology 140, 55.

[27] Galton, F. (2005). *Finger prints*. Mineola, NY: Dover Publications. (Original work published 1892)

[28] Jain, A., *Pores and Ridges: High-Resolution Fingerprint Matching Using Level 3 Features*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 1, January 2007.

[29] Orandi, S., Libert, J. M., Grantham, J. D., Ko, K., Wood, S. S., Wu, J. C. Effects of JPEG 2000 Image Compression on 1000 ppi Fingerprint Imagery, NIST Interagency Report 7778 (NISTIR7778), National Institute of Standards and Technology, Gaithersburg, MD, May 2011. http://www.nist.gov/manuscript-publication-search.cfm?pub_id=908204 (Retrieved 05/01/2013).

[30] Orandi, S., Libert, J. M., Grantham, J. D., Lepley, M., Bandini, B., Ko, K., Petersen, L. M., Wood, S. S., Harvey, S. G. Examination of Downsampling Strategies for Converting 1000 ppi Fingerprint Imagery to 500 ppi. NIST Interagency Report 7839 (NISTIR7839), National Institute of Standards and Technology, Gaithersburg, MD, January 2013. http://nvlpubs.nist.gov/nistpubs/ir/2013/NIST.IR.7839.pdf (Retrieved 05/01/2013).

[31] Conover, W. J. Practical Nonparametric Statistics, Third Edition, Wiley, 1999, pp. 367-373.

[32] Hollander, M., and D. A. Wolfe. Nonparametric Statistical Methods. Hoboken, NJ: John Wiley & Sons, Inc., 1999.

[33] Wilcoxon, Frank . *Individual comparisons by ranking methods*. Biometrics Bulletin 1 (6), (1945), 80–83. http://www.jstor.org/sici?sici=0099-4987%28194512%291%3A6%3C80%3AICBRM%3E2.0.CO%3B2-P (Retrieved 05/01/2013)

[34] Wright, P, S. *P-values for simultaneous inference*. Biometrics, Vol. 48, No. 4 (1992), pp. 1005-1013.

[35] "WSQ Grayscale Fingerprint Image Compression Specification" Version 3.1. Https://www.fbibiospecs.org/docs/WSQ_Grayscale_Specification_Version_3_1.pdf. Retrieved 2010-01-11.

[36] Bradley, Jonathan N., Brislawn, Christopher M., Hopper, Thomas. (1993). "FBI wavelet/scalar quantization standard for grayscale fingerprint image compression", SPIE Conference on Visual Information Processing II, eds. Huck, Friedrich O., Juday, Richard D. pp. 293-304.

[37] Bradley, J. N., Brislawn, C. M. (1994). "The wavelet/scalar quantization compression standard for digital fingerprint images". IEEE International Symposium on Circuits and Systems, 1994. ISCAS '94., 1994. Vol 3. pp. 205-208.

[38] Brislawn, Christopher M. (1996). "Wavelet Scalar Quantization Compression Standard for Fingerprint Images". Proceeding Conference on Signal Image Processing and Applications.

[39] Brislawn, Christopher M. (2002). The FBI Fingerprint Image Compression Standard. http://wwwc3.lanl.gov/~brislawn/FBI/FBI.html, updated 25 June 2001 (accessed 01/31/2014).