EVALUATION OF IMAGE QUALITY FEATURES VIA MONOTONIC ANALYSIS*

Lance M. Kaplan
US Army Research Laboratory
Adelphi, MD 20783

Rick S. Blum
Lehigh University
Bethlehem, PA 18015

ABSTRACT

This work introduces novel monotonic analysis to determine whether or not proposed image quality (IQ) measures are consistent with human measured perceptual quality scores. Specifically, the analysis performs a generalized likelihood ratio test over the H_1 hypothesis that the IQ measures and the corresponding perceptual measurements are related via a monotonic function versus the null hypothesis that the functional relationship is arbitrary. This paper evaluates six proposed IQ measures against mean opinion scores using the new monotonic analysis.

1 INTRODUCTION

The next generation of night vision goggles and night scopes will fuse image intensified (I2) and long wave infrared (LWIR) to create a hybrid image that will enable soldiers to better interpret their surroundings during nighttime missions. The key to such systems is the determination of the best image fusion algorithm for a specific task. A number of image fusion algorithms have been proposed in the literature, e.g. (Zhang and Blum 1999). Currently, a scientific evaluation of such algorithms requires extensive and expensive human perception studies to determine how well soldiers can perform a specific task. What is needed is an image quality (IQ) measure than can automatically quantify the utility of image fusion algorithms.

The ultimate goal is an image model that is able to predict human performance given a few IQ measures as input parameters. This papers demonstrates the monotonic correlation as a tool to score the myriad of measures based upon how well an arbitrary monotonic curve is able to fit the relationship between computed IQ features and human performance. Previous work investigated the monotonic correlation for the human task of classification (Kaplan et al. 2008b) when fusing the I2 and LWIR bands. This paper uses the experimental results from (Chen and Blum 2008) which also considers the fusion of the I2 and LWIR bands. In that work, multiple humans score imagery resulting from 6 common fusion algorithms based on perceived perceptual quality over 28 different scenes (a total of 168 images). Furthermore, 6 full-reference IQ measures were calculated over the 168 images.

This paper is organized as follow. Section 2 details the perceptual experiments including the image fusion algorithms, IQ evaluation, and human perceptual scoring that was used. Then, Section 3 introduces the tools for monotonic analysis. These tools are used to evaluate proposed IQ measures in Section 4. Finally, Section 5 provides concluding remarks.

2 Perceptual Experiment

The perceptual experiment consisted of applying image fusion algorithms over registered I2 and LWIR images, calculating various IQ measures and measuring human preference over the fused images. The details are provided in the following subsections.

2.1 Image Fusion Algorithms

The image fusion algorithm takes input from a number of source images and generates a single fused image that is presented to the human user for interpretation. Image fusion has a number of applications including remote sensing, concealed weapon detection, and night vision (Simone et al. 2002; Chen et al. 2005; Blum and Liu 2006). Two main classes of fusion algorithms exist. The first generates a gray scale image by determining which information to include from the various source images (see (Zhang and Blum 1999) for an excellent review of such methods). The second class generates a color image by mapping different source images into different color spaces, e.g., (Waxman et al. 1997). This class of methods is only appropriate when three or fewer sources are used. This work only considers gray scale fusion.

For the experiments, six gray scale image fusion meth-

*Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-06-2-0020. The views and conclusions contained in this document are those of the authors and should not be interpreted as the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 DEC 2008	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Evaluation Of Image Quality Features Via Monotonic Analysis	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army Research Laboratory Adelphi, MD 20783**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002187. Proceedings of the Army Science Conference (26th) Held in Orlando, Florida on 1-4 December 2008, The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	

ods were implemented including: 1) simple pixel averaging (Bender et al. 2003), 2) discrete wavelet transform (DWT) (Huntsberger and Jawerth 1993), 2) Filter-Subtract-Decimate pyramid (FSD) (Anderson 1988), 3) Laplacian pyramid (LAP) (Burt and Adelson 1983), 5) Morphological pyramid (Morph) (Toet 1989), and 6) Shift invariant DWT (SiDWT) (Rockinger 1997). Note that only the fused and not the original I2 and LWIR channels were evaluated in the human perception experiments. The details about the exact implementation of the fusion methods is available in (Chen and Blum 2008). Figure 1 shows examples of imagery generated by these six fusion methods.

2.2 Image Quality Measures

A full reference IQ measure quantifies the similarity of a processed image against the original image (or images). A number of such IQ measures have been proposed to evaluate image compression algorithms. The typical metrics to evaluate compressed imagery, e.g., mean squared error (mse) and peak signal to noise ratio (PSNR) are known to be poor IQ metrics, and more relevant metrics are described and evaluated in (Wang et al. 2004) for the application of image compression. These measures are easily adapted for image fusion algorithms where the fused IQ is the weighted average of the IQ measure between the fused image and each of the source images (Piella 2004; Xydeas and Petrović 2000; Qu et al. 2002; Chen and Varshney 2007; Wang et al. 2004; Chen and Blum 2008). In effect, these full reference features quantify how well "salient features" in the fused imagery matches the "salient features" in the source images. Table 1 summarizes the six potential full reference IQ measures considered in this work and point to appropriate references.

2.3 Perceptual Image Evaluation

This paper performed perceptual evaluation of 28 scenes consisting of co-registered I2 and LWIR imagery. Specifically, these scenes were processed using the 6 fusion algorithms described in Section 2.1. Figure 1 shows an example of one of the scenes. The source I2 and LWIR images are provided in Figures 1(a) and (b), respectively. The I2 image provides finer resolution, texture and better context than the LWIR image. On the other hand, the contrast is better in the LWIR image. In fact, the human is only visible in the LWIR image. Figures 1(c)-(h) provide the corresponding fused images. All the fused images contain the signature of the human. The pixel-averaging provided the poorest contrast. In terms of contrast and texture, one could argue that the DWT and Morph images are slightly better than the FSD, LAP, and SiDWT images.

Human observers provided opinion scores ranging from one (worst quality) to 10 (best quality) for each of the 168 fused images. The mean opinion scores (MOS) and associated sample variances are provided in (Chen and Blum 2008) for each fusion method and each scene. In this work, the MOS represents the perceptual score of the images.

3 Monotonic Analysis

A potential IQ measure is simply a deterministic mapping of an image into a scaler that quantifies how well the image actually portrays the scene. For image fusion applications, the IQ measure indicates how well the relevant details in the source images are preserved in the fused image. On the other hand, the perceptual scores can be viewed as noisy measurements. A repeat of the perception experiments with the same imagery should lead to similar but not the same results. Thus, it should be reasonable to model the perceptual scores as the nominal result embedded in noise.

The actual individual preference for a given image can be biased by the content in the scene. In other words, it is possible for one fusion method to generate a desirable image for one scene, but not the other. As a result, the relative rankings of the utility of the different image fusion algorithms can change from scene to scene. A desirable IQ measure should track the relative rankings over the various scenes. This section details a hypothesis test to determine whether or not the proposed IQ measure and the perception results demonstrate a consistent monotonic relationship over the scenes under test. First, the data models for the IQ and perceptions results are provided. Next, the concept of monotonic correlation is introduced. Finally, the relationship between the monotonic correlation and a generalized likelihood ratio test is shown.

3.1 Data Models

Given that N_f fusion algorithms are under consideration, let the $N_f \times 1$ vector $\mathbf{x}$ represent a given IQ measure evaluated over the N_f fused images associated to a given scene. Likewise, let the $N_f \times 1$ vector $\mathbf{y}$ represent the MOS values collected over the same N_f fused images. The pair (x_i, y_i) represents the IQ measure and MOS value for the i-th fused image. The vector $\mathbf{x}$ is deterministic because it represents IQ results. On the other hand, we model the MOS value for the i-th fusion method as

$$y_i = \mu_i + n_i, \tag{1}$$

where $n_i \sim N(0, \sigma_n^2)$ due to the central limit theorem. The mean value μ_i is taken to be the sample mean of the opinion scores tabulated over the i-th fused image. The variance of the measurement noise σ_n^2 is taken to be the sample variance over all opinion scores for the scene divided by N_f.

A statistic to evaluate the usefulness of the proposed IQ measure that generates the vector $\mathbf{x}$ must quantify how well the pairs $(\mathbf{x}, \mathbf{y})$ support the hypothesis that there exist an arbitrary monotonic function $h_{\text{mono}}(\cdot)$ such that $\mu_i = h_{\text{mono}}(x_i)$. Equivalently, the monotonic hypothesis indicates that either $x_i > x_k$ implies $\mu_i \geq \mu_k$ (monotonically in-

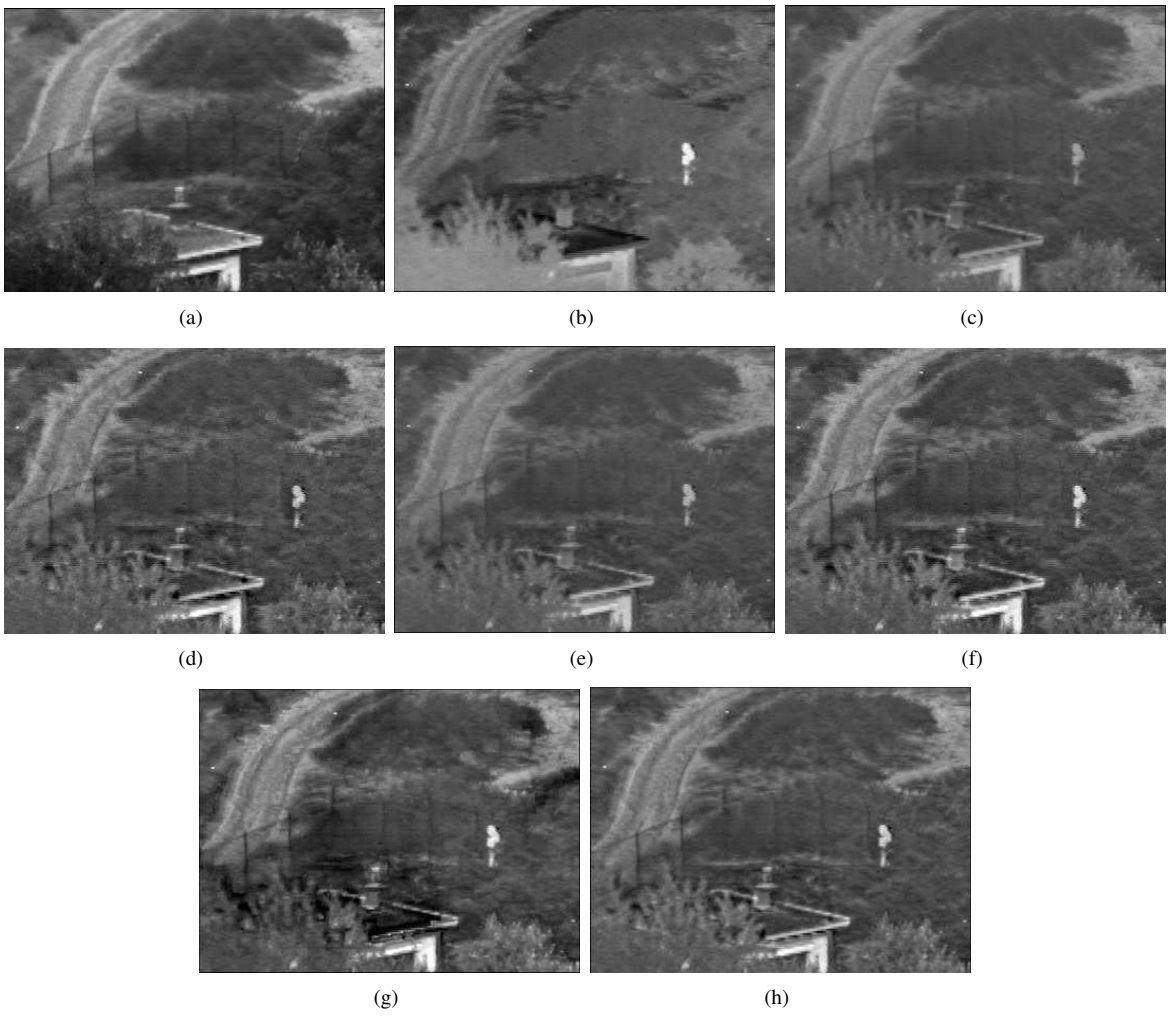


Figure 1: Example of source and fused images: (a) I2, (b) LWIR, (c) pixel averaging, (d) DWT, (e) FSD, (f) LAP, (g) Morph, (h) SiDWT.

1	Universal Quality Index (UI)	Average Structure SIMilarity (SSIM) index between fused and reference images	(Wang et al. 2004)
2	Information Measures (MI)	Average mutual information between fused and reference images (bin size = 16)	(Qu et al. 2002)
3	Objective Measure (QE)	Average objective edge information between fused and reference images	(Xydeas and Petrović 2000)
4	Mannos Quality Index (Q_M)	HVS quality index using the Mannos & Sakrison constrast sensitivity filter	(Chen and Blum 2008)
5	Barton Quality Index (Q_B)	HVS quality index using using the Barton contrast sensitivity fitler	(Chen and Blum 2008)
6	Difference Quality Index (Q_D)	HVS quality index using using the difference of Gaussian contrast sensitivity filter	(Chen and Blum 2008)

Table 1: List of potential full-reference IQ measures evaluated in this paper.

creasing) or $x_i > x_k$ implies $\mu_i \leq \mu_k$ (monotonically decreasing). In reality, a proper IQ measure should exhibit a monotonically increasing relationship with human performance. However, if the relationship is monotonically decreasing, the proposed feature can trivially be transformed into a proper IQ measure via a negative or reciprocal operation. Because $\mathbf{x}$ and $\mathbf{y}$ are the input and noisy output values to the function $h_{\text{mono}}(\cdot)$, respectively, we refer to $\mathbf{x}$ and $\mathbf{y}$ as the input and output vectors, respectively, in the sequel.

3.2 Monotonic Correlation

The standard Pearson correlation can be viewed as the square root of the coefficient of determination (R^2) that is obtained by fitting a line to the samples (x_i, y_i) for $i = 1, \ldots, N_f$. Motivated by this interpretation of the Pearson correlation, we define the *monotonic correlation* (MC) as the R^2 value that is obtained by fitting an arbitrary monotonic curve to the (x_i, y_i) samples. To this end, the samples are reindexed so that that the values of $\mathbf{x}$ are in ascending order, i.e., $x_1 \leq x_2 \leq \cdots \leq x_{N_f}$. Then, the monotonic fit is determined by selecting values $\hat{\mathbf{y}}$ that are in either ascending or descending order such that means squared difference between $\hat{\mathbf{y}}$ and $\mathbf{y}$ is minimized. The monotonic fit can be found by solving two Quadratic Programming (QP) problems

$$\begin{aligned} &\hat{\mathbf{y}}_{\uparrow} = \arg\min \|\mathbf{y} - \mathbf{z}\|^2, \quad \hat{\mathbf{y}}_{\downarrow} = \arg\min \|\mathbf{y} - \mathbf{z}\|^2 \\ &\text{s.t. } z_1 \leq z_2 \leq \ldots \leq z_N, \quad \text{s.t. } z_1 \geq z_2 \geq \ldots \geq z_N, \end{aligned} \tag{2}$$

Note that for the case that some input values are equal, e.g., $x_i = x_{i+1} = \cdots x_{i+k}$, then the corresponding inequalities constraints become active, i.e, $z_i = z_{i+1} = \cdots = z_{i+k}$, because the arbitrary monotonic function cannot produce more that one output value for the same input value. Then, $\hat{\mathbf{y}}$ is the $\hat{\mathbf{y}}_{\uparrow}$ or $\hat{\mathbf{y}}_{\downarrow}$ that leads to the lowest residual error,

$$\hat{\mathbf{y}} = \begin{cases} \hat{\mathbf{y}}_{\uparrow} & \text{if } \|\mathbf{y} - \hat{\mathbf{y}}_{\uparrow}\|_2 < \|\mathbf{y} - \hat{\mathbf{y}}_{\downarrow}\|_2, \\ \hat{\mathbf{y}}_{\downarrow} & \text{otherwise.} \end{cases} \tag{3}$$

Finally, the R^2 value for the monotonic fit determines the MC

$$\rho_{\text{mono}} = \pm\sqrt{1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\sigma_s^2}} \tag{4}$$

where σ_s^2 is the sample variance of the values in $\mathbf{y}$ scaled by N_f, and the sign is positive by convention if $\hat{\mathbf{y}}$ is ascending, i.e., $\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\uparrow}$, and negative otherwise. Alternatively, the MC can be computed via the Pearson correlation of $\mathbf{y}$ and $\hat{\mathbf{y}}$. For purposes of integrating likelihoods ratios over disparate scenes (see next subsection), we define $\rho_{\uparrow}$ (*isotonic increasing correlation*) or $\rho_{\downarrow}$ (*isotonic decreasing correlation*) by substituting $\hat{\mathbf{y}}_{\uparrow}$ or $\hat{\mathbf{y}}_{\downarrow}$, respectively, for $\hat{\mathbf{y}}$.

The heart of calculating ρ_{mono} is solving the two QP problems in (2). Because the function to minimize is convex and the feasible region defined by the constraints is convex, there is a unique minima. Therefore, the optimal solution can be found without worrying about the initial guess for $\hat{\mathbf{y}}$. In fact, these QP problems are examples of the same well known isotonic regression problem, and the pool adjacent violators (PAV) algorithm can determine the exact optimal values of $\hat{\mathbf{y}}_{\uparrow}$ and $\hat{\mathbf{y}}_{\downarrow}$ in N_f steps, (Barlow et al. 1972; Hanson et al. 1973). In fact, it is shown in (Best and Chakravarti 1990; Pardalos and Xue 1999) that an efficient coding of the PAV requires only O(N_f) operations.

Note that the PAV does not account for the active constraints. To force the active constraints when $x_i = x_{i+1} = \cdots = x_{i+k}$, the output values corresponding to equal input values are replaced by the corresponding mean value, e.g., $y_j \leftarrow \frac{1}{k+1}\sum_{n=i}^{i+k} y_n$ for $j = i, \ldots, i+k$ before entering the PAV. As shown in (Kaplan et al. 2008a), this modified PAV will produce the optimal results.

The MC possesses many interesting properties. Like linear correlation, it is invariant to linear transformation of the input and output sequences. It is also invariant to any monotonic transformation of the input sequence, because such a transformation does not change the ordering of the elements to solve (2). The MC is **not** invariant to monotonic transformation of the output sequence. The calculation of the model error places a higher penalty when the miss-ordered values in the output sequence have higher variance than when these values are tightly clustered together. As a result, the MC is lower when ordering the input leads to larger non-monotonic "swings" in the output sequence (see Figure 2). Finally, it not difficult to show that $|\rho_{\text{lin}}| \leq |\rho_{\text{mono}}|$ because any linear function is monotonic and the monotonic fit will be at least as good as the linear fit. Figure 3 show examples of linear and monotonic fits to a scatter plot of points (x_i, y_i) and the corresponding correlation values. The figure also demonstrates the fit of a logistic function to the data. While the logistic function provides a better fit than linear, the logistic function does not provide a good fit for the points whose x value is grater thane 0.95. This is due to the fact that the logistic function can not model two or more inflection points.

3.3 Hypothesis Test

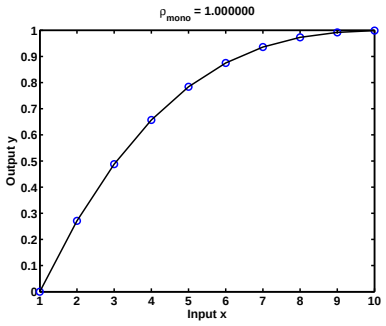
This section connects the correlation analysis of the previous section to a hypothesis test. The null H_0 hypothesis is that the IQ feature is not monotonically related to human performance, and the H_1 hypothesis is that the monotonic relationship does exist. Under the null hypothesis, the ground truth human performance is related to the actual IQ feature via an arbitrary function $h(\cdot)$ so that based on (1),

$$y_i = h(x_i) + n_i, \tag{5}$$

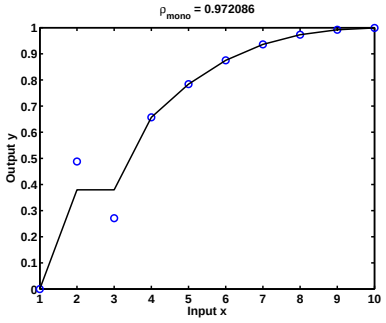
Likewise, for the H_1 hypothesis,

$$y_i = h_{\text{mono}}(x_i) + n_i. \tag{6}$$

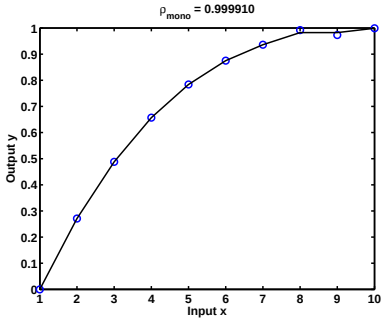
It is well known that comparing the likelihood ratio be-



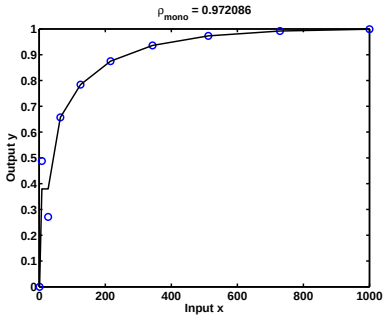
(a)



(b)

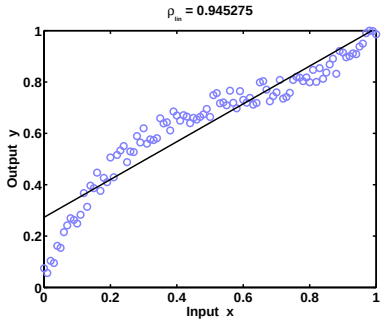


(c)

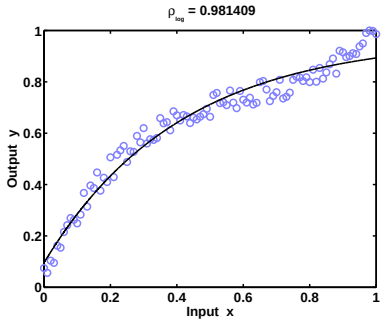


(d)

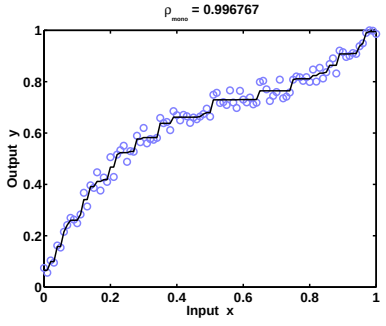
Figure 2: Examples of monotonic fit and MC: (a) Perfect fit ($\rho_{\text{mono}} = 1.0000$), (b) a single miss-ordering between two input features causing a large output "swing" lowers the correlation to $\rho_{\text{mono}} = 0.9721$, (c) a single miss-ordering between two input features causing a small output "swing" only lowers the correlation to $\rho_{\text{mono}} = 0.9999$, and (d) a cubic stretching of the feature values in (b) does not change the fit or MC.



(a)



(b)



(c)

Figure 3: Examples of curve fits and correlation values: (a) linear, (b) logistic, and (c) monotonic.

tween two simple[1] hypotheses to a threshold leads to the universally most powerful (UMP) test. From (5), the likelihood for the null hypothesis is

$$f(\mathbf{x}|H_0) = \frac{1}{(2\pi\sigma_n^2)^{\frac{N_f}{2}}} \exp\left(-\frac{\sum_{i=1}^{N_f}(y_i - h(x_i))^2}{2\sigma_n^2}\right), \tag{7}$$

and from (6), the likelihood for H_1 is

$$f(\mathbf{x}|H_1) = \frac{1}{(2\pi\sigma_n^2)^{\frac{N_f}{2}}} \exp\left(-\frac{\sum_{i=1}^{N_f}(y_i - h_{\text{mono}}(x_i))^2}{2\sigma_n^2}\right). \tag{8}$$

The nonlinear functions $h_{\text{mono}}(\cdot)$ and $h(\cdot)$ are unknown *a priori*, and it is not possible to compute the likelihood ratio. Therefore, we resort to the generalized likelihood ratio test (GLRT) where the unknown parameters in (7)-(8) are replaced my their ML estimates. Actually, when estimating the ML estimates of the non-linear functions, one only needs to consider the function values at x_i, i.e., $g_i = h(x_i)$ (or $g_i = h_{\text{mono}}(x_i)$) under the H_0 (or H_1) hypothesis, for $i = 1, \ldots, N_f$. The maximization of (8) is equivalent to the monotonic regression given by (2) and (3). On the other hand, the maximization of (7) trivially selects $h(\cdot)$ so that $g_i = y_i$ for $i = 1, \ldots, N_f$. Therefore the generalized likelihood ratio (GLR) is

$$\Lambda = \exp\left(-\frac{\sum_{i=1}^{N_f}(y_i - \hat{y}_i)^2}{2\sigma_n^2}\right), \tag{9}$$

and given (4), the relationship between the GLR and the MC is derived to be

$$\Lambda = \exp\left((\rho_{\text{mono}}^2 - 1)\frac{\sigma_s^2}{2\sigma_n^2}\right). \tag{10}$$

Note that σ_s^2 represents the inter-fusion method spread of the perception scores. In a similar vain, σ_n^2 represents the intra-fusion method spread of the perception scores, and the ratio

$$K = \frac{\sigma_s^2}{\sigma_n^2} \tag{11}$$

is analogous to a class separability criteria used in discriminant analysis (Fukunaga 1990). We refer to this ratio as the separability ratio in the sequel. Since the magnitude of the correlation is bound by zero and one, Λ can take on values from $\exp\left(-\frac{1}{2}K\right)$ to one. The GLR is always less than or equal to one because the arbitrary H_0 fit can never be worse than the monotonic fit. GLR values close to one indicate a high likelihood that the fit between the $\mathbf{x}$ and $\mathbf{y}$ is monotonic.

When the likelihood ratio is much greater than one, there is compelling evidence that the H_1 hypothesis is true. In other words, the support for the H_1 hypothesis is statistically significant. Because a high MC leads to a GLR value close to one no matter the separability ratio, it is difficult to determine the statistical significance of the large value. Actually, the significance of the MC is intertwined with the spread of possible GLR values under the null hypothesis. Note that the input $\mathbf{x}$ influences the MC strictly by how it sorts out the output $\mathbf{y}$. Under the null hypothesis, the ranking of perception results are unrelated to the ranking of the feature values. Therefore, the GLR value could result from any arbitrary ordering of the perception values, i.e., the index i for y_i is arbitrary. Thus, to gain insight about the significance of the GLR value (based on a particular $\mathbf{x}$), it is instructive to take the ratio of Λ over the expected value of Λ given random input feature values drawn from an uninformative prior, which we label as $\bar{\Lambda}$. Under an uninformative prior for the input $\mathbf{x}$, the sorting of $\mathbf{y}$ is one of $N_f!$ possibilities with equal probability. For small N_f, $\bar{\Lambda}$ can be computed by averaging over all $N_f!$ possible values of Λ, but when N_f is large, one must resort to averaging over Monte-Carlo trials. We define the normalized GLR as

$$\tilde{\Lambda} = \frac{\Lambda}{\bar{\Lambda}} \tag{12}$$

so that $\tilde{\Lambda}$ can exceed one. When the separability ratio is $K = 0$, the normalized GLR is always one, $\tilde{\Lambda} = 1$, and when the MC is close to one, there is no compelling evidence to support the H_1 hypothesis. When $|\rho_{\text{mono}}| = 1$ and $K > 0$, $\tilde{\Lambda} > 1$. As K becomes larger, so does $\tilde{\Lambda} > 1$, and the evidence to support the H_1 hypothesis becomes more significant. For values of $|\rho_{\text{mono}}|$ near one, $\tilde{\Lambda}$ increases as K becomes larger than zero. However, as K approaches infinity, $\tilde{\Lambda}$ goes down to zero. In other words, when the spread is zero, the results are meaningless to make any conclusions. As the spread goes to infinity, there is no measurement error and $(\mathbf{x}, \mathbf{y})$ must trace out a monotonic curve. In between, a $|\rho_{\text{mono}}|$ near one may be significant.

When performing monotonic analysis over multiple scenes, the sign of the correlation should be consistent from one scene to the next. Otherwise, it is impossible to determine if a higher feature score translates to high or low quality. As a result, one should consider the *isotonically increasing GLR* and *isotonically decreasing GLR* by substituting $\rho_\uparrow$ or $\rho_\downarrow$, respectively, for ρ_{mono} in (10). Then if $\tilde{\Lambda}_{\uparrow,s}$ and $\tilde{\Lambda}_{\downarrow,s}$ are the normalized isotonic likelihoods for the s-th scene, the overall normalized GLR for all N_s scenes is

$$\tilde{\Lambda} = \max\left(\prod_{s=1}^{N_s}\tilde{\Lambda}_{\uparrow,s}, \prod_{s=1}^{N_s}\tilde{\Lambda}_{\downarrow,s}\right). \tag{13}$$

4 Data Analysis

The monotonic analysis described in Section 3 was used to evaluate the results of the perceptual experiment described in Section 2. Table 2 summarizes statistics about ρ_{mono} and

[1] A simple hypothesis is one with completely known likelihood function which has no unknown parameters.

Statistic	Q_M	Q_B	Q_D	QE	UI	MI
$\max \lvert\rho_{\text{mono}}\rvert$	0.999	1.000	1.000	1.000	1.000	0.999
$\min \lvert\rho_{\text{mono}}\rvert$	0.369	0.466	0.687	0.762	0.375	0.598
mean ρ_{mono}	0.586	0.662	0.955	0.943	0.572	-0.861
# $\rho_{\text{mono}} > 0$	23	25	28	28	24	0
# $\tilde{\Lambda}_s > 4$	13	9	20	15	2	6
# $\tilde{\Lambda}_s > 1$	18	11	24	24	6	16
$\tilde{\Lambda}$	4.282e-22	5.227e-24	4.924e+22	5.833e+19	8.264e-44	7.221e-03

Table 2: Statistics describing the significance of the perception results via the monotonic analysis.

$\tilde{\Lambda}_s$ over each of the scenes. For every IQ measure, there is at least one scene where the monotonic fit between the measure values and the MOS is very good. However, on average the monotonic fit is only high for the Q_D and QE IQ measures. Furthermore, the MCs for the Q_D and QE measures are positive for all scenes. Likewise, the MCs for the MI measure is always negative for the MI measure. When the normalized GLR threshold is four, the Q_D measure is significant for the most scenes followed by QE. In fact, for all but four scenes, the normalized GLRs for Q_D and QE exceeds one. Finally, the overall normalized GLR $\tilde{\Lambda}$ was computed by (13) and is included in Table 2. Overall, the 28 scenes support the fact that Q_D and QE exhibit a monotonic relationship with human perceptual quality. The monotonic analysis does not support the other four measures as good IQ measures. The relative rankings of the measures via the overall normalized GLR is consistent with the scoring mechanisms presented in (Chen and Blum 2008) with the exception that MI is third as opposed to six. This is due to the fact that the analysis in this paper accepts monotonic decreasing relationships since the measure can be transformed into an IQ measure by taking the reciprocal. Nevertheless, the overall normalized GLR score for MI is well below a value of one. Finally, none of the feature scoring mechanisms in (Chen and Blum 2008) indicate the extent of the support for the hypothesis that a proposed feature is a good IQ measure.

5 Conclusions

This work provides novel analysis to measure the suitability of proposed IQ measures using results from human perceptual experiments. The basis of this foundation is the use of a MC statistic that determines to what degree does a monotonic relationship exists between a proposed IQ measure and human perceptual score. As demonstrated in this paper, the MC is more general than linear and logistic correlations. This work also shows the connection between the MCs and a hypothesis test attempting to decide if the proposed IQ measures exhibit a monotonic relationship with human perception performance. Finally, the paper introduces the concept of the normalized GLR to evaluate the statistical significance of the MC, or corresponding GLR value, in light of any random ordering of the human perception results. The monotonic analysis was used to evaluate 6 IQ measures. The analysis reveals the effectiveness of the objective measure (QE) and the HVS quality index using the difference of Gaussian contrast sensitivity filter (Q_D).

Acknowledgements

The authors are extremely grateful to Dr. Yin Chen of Lehigh University for providing Matlab code to implement many of the full-reference measures.

References

Anderson, C. H.: 1988, Filter-subtract-decimate hierarchical pyramid signal analyzing and aynthesizing technique. US Patent 4,718,104.

Barlow, R., D. Barholomew, J. Bremner, and H. Brunk: 1972, *Statistical Infernece under Order Restrictions*. Wiley, New York.

Bender, E. J., C. E. Reese, and G. S. van der Wal: 2003, Comparison of additive image fusion vs. feature-level image fusion techniques for enhanced night driving. *Proc. of SPIE*, volume 4796, 140–151.

Best, M. and N. Chakravarti: 1990, Active set algorithms for isotonic regression: A unifying approach. *Mathematical Programming*, **47**, 425–439.

Blum, R. S. and Z. Liu, eds.: 2006, *Multi-Sensor Image Fusion and Its Applications*. CRC Press, Boca Raton, FL.

Burt, P. J. and E. H. Adelson: 1983, The Laplacian Pyramid as a compact image code. *IEEE Transactions on Communications*, **COM-31,4**, 532–540.

Chen, H. and P. K. Varshney: 2007, A human perception inspired quality metric for image fusion based on regional information. *Image Fusion*, **8**, 193–207.

Chen, H.-M., S. Lee, R. M. Rao, M.-A. Slamani, and P. K. Varshney: 2005, Imaging for concealed weapon detection: A tutorial overview of development in imaging sensors and processing. *IEEE Signal Processing Magazine*, **22**, 52–61.

Chen, Y. and R. S. Blum: 2008, A new automated quality assessment algorithm for image fusion, to appear in *Image and Vision Computing*.

Fukunaga, K.: 1990, *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 2nd edition.

Hanson, D., G. Pledger, and G. Wright: 1973, On consistency in monotonic regression. *The Annals of Statistics*, **1**, 401–421.

Huntsberger, T. L. and B. D. Jawerth: 1993, Wavelet-based sensor fusion. *Proc. of SPIE*, volume 2059, 488–498.

Kaplan, L. M., S. D. Burks, R. S. Blum, R. K. Moore, and Q. Nguyen: 2008a, Analysis of image quality for image fusion via monotonic correlation, sumitted for pulication in *IEEE Journal of Special Topic in Signal Processing*.

Kaplan, L. M., S. D. Burks, R. K. Moore, and Q. Nguyen: 2008b, Monotonic correlation analysis of image quality measures for image fusion. *Proc. of SPIE*, volume 6941.

Pardalos, P. and G. Xue: 1999, Algorithms for a class of isotonic regression problems. *Algorithmica*, **23**, 211–222.

Piella, G.: 2004, New quality measures for image fusion. *Proc. of the 7th Intl. Conf. on Information Fusion*, Stockholm, Sweden, 542–546.

Qu, G., D. Zhang, and P. Yan: 2002, Information measure for performance of image fusion. *Electronic Letters*, **38**, 313–315.

Rockinger, O.: 1997, Image sequence fusion using a shift invariant wavelet transform. *Proc. of the IEEE Intl. Conf. on Image Processing (ICIP)*, volume 3, 288–291.

Simone, G., A. Farina, F. C. Morabito, S. B. Serpico, and L. Bruzzone: 2002, Image fusion techniques for remote sensing applications. *Image Fusion*, **3**, 3–15.

Toet, A.: 1989, A morphological pyramidal image decomposition. *Pattern recognition letters*, **9**, 255–261.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simincelli: 2004, Image quality assessment: From error measurement to structural similarity. *IEEE Trans. on Image Processing*, **13**.

Waxman, A. M., A. N. Gove, D. A. Fay, J. P. Racamato, J. E. Carrick, M. C. Seibert, and E. D. Savoye: 1997, Color night vision: Opponent processing in the fusion of visible and IR imagery. *Neural Networks*, **10**, 1–6.

Xydeas, C. and V. Petrović: 2000, Objective image fusion performance measure. *Electronic Letters*, **36**, 308–309.

Zhang, Z. and R. S. Blum: 1999, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. *Proceedings of the IEEE*, **87**, 1315–1326.