Image Fusion for Human Observers: How Should We Choose the Method?

Murray Loew, James Bonick, Clarence Walters
US Army RDECOM CERDEC
Night Vision & Electronics Sensors Directorate
Ft. Belvoir, VA 22060

Image fusion is used to improve target detection and identification. In human-observer applications it is useful to rank fusion methods according to how well they assist the observer in a decision task.

Two images (medium- and long-wave infrared), acquired for each of a number of outdoor scenes, were fused by each of nine methods. For each scene, a set of observers assessed each of the 36 pairwise combinations of fused images, choosing from each pair the one that was deemed best for target identification. We used that set of preferences to rank the fusion methods for their effectiveness in the identification task.

A classical technique for ranking these "discriminal processes" is Thurstone's Law of Comparative Judgment and its implementation as the Thurstone-Mosteller (TM) Method of Paired Comparisons, which is reviewed briefly here. To make meaningful statements about preferences, one should have a measure of uncertainty for each rank. The TM method, however, cannot readily provide such a measure. An alternative, the Bradley-Terry (BT) method, does permit calculation of confidence intervals for ranks. To our knowledge, BT has not previously been applied in the evaluation of fusion methods.

We present results from a multi-observer, multi-view trial, evaluated using TM and BT. The methods yield similar rankings of the fusion methods. But the additional information provided by BT – that is, whether there are significant differences between the ranks – can have a substantial impact on the implementation of fusion in real systems. There could be meaningful tradeoffs among fusion methods – e.g., performance vs. computation time – that may not be exploited in the absence of those insights.

1 Introduction

1.1 Background

Image fusion is used widely in military applications to improve target detection and identification. In general, the method chosen for fusion will depend on whether the fused imagery is to be interpreted by a human observer or processed by a machine. In the former case, therefore, it is useful to rank fusion methods according to the preferences expressed by human observers. This work, part of a continuing effort, applies and compares two ranking methods to data from an observer experiment, and identifies distinct advantages of one. The present study was motivated by prior efforts in this area [1-3], which demonstrated some of the underlying principles, but did not make the differences explicit.

We distinguish this work from other related efforts that evaluated fusion methods on the basis of observers' performances on a specific task (e.g., target identification) [3]. In this work we asked observers for only their preference with respect to a task ("Which do you prefer for target identification?") within a pair of images; they did not actually perform the task.

In comparing the various fusion methods, we are not simply varying a single continuous parameter. We wish to create an interval scale that allows comparison of the various "stimuli" – the set of images created by a variety of fusion methods.

Thurstone introduced an approach to this problem [12]. He postulated a series of stimuli to which the subject can respond differentially (the "discriminal process") with respect to some attribute. The task is to locate these on a continuum so that we can account for the responses given by the observer. A given stimulus does not always excite the same discriminal process, and so if a given stimulus is presented to an observer a number of times, we can think of a frequency distribution of the psychological continuum of discriminal processes associated with that stimulus [14]. This led to Thurstone's Method of Paired Comparisons [12] (also known as the Law of Comparative Judgment), which comprised five cases, ranging from complete generality to a set of restrictive assumptions. In all cases, the observer examines a pair and reports a preference between stimulus i (fusion method i, in our setting) and stimulus j; either $i > j$ or $i < j$. The preference data are compiled into matrices; each preference matrix contains in each cell c(i,j) the number of times, for a given scene, that fusion method j was preferred to fusion method i, aggregated across all observers.

1.2 Ranking and Statistics

Mosteller [11] stated Thurstone's general conditions as follows:

(1) There is a set of stimuli which can be located on a subjective continuum (a sensation scale, usually not having a measurable physical characteristic).
(2) Each stimulus when presented to an individual gives rise to a sensation in the individual.
(3) The distribution of sensations from a particular stimulus for a population of individuals is normal (Gaussian).
(4) Stimuli are presented in pairs to an individual, thus giving rise to a sensation for each stimulus. The individual compares these sensations and reports which is greater.
(5) It is possible for these paired sensations to be correlated.
(6) Our task is to space the stimuli (the sensation means) except for a linear transformation.

Thurstone's Case V assumes equal standard deviations of sensations corresponding to stimuli, and zero correlations between pairs of sensations. Mosteller [11] examined and extended the analysis of Case V and showed that the assumption of zero correlations can be relaxed to an assumption of equal correlations between pairs, with no change in method. He also showed how to estimate the stimulus positions on the sensation scale using a least-squares technique. He did not, however, present a method for accounting for imperfect data; that is, how to draw inferences from the data recorded in the observer study, which are estimates of the preferences.

To be of real use, methods for ranking that are based on paired comparisons should provide a measure of their uncertainties – confidence intervals – which will permit tests of significance. This will help answer several questions: (1) is any algorithm's performance exceptional? (2) given the high correlation of the bands, is image fusion a viable and effective technique when applied to imagery from the two modalities?

If X_i and X_j are single sensations evoked in an individual by stimuli Oi and Oj, then we let the mean of X_i be S_i, with corresponding variance σ_i, and correlation ρ_{ij} between two sensations. The preference matrix (defined in Sec. 1.1) provides the probabilities $P(X_i > X_j)$. What we wish to compute are the locations S_i and S_j of the sensations (responses). To compute the locations from the probabilities is, in the general case, impossible, because with n stimuli, we have only n(n-1)/2 observer equations, and there are n scale values, n standard deviations, and n(n-1)/2 correlations, all of which are unknown.

As Thurstone and others have done, we simplify (use his Case V), and set $\rho_{ij} = 0$. This permits a solution (e.g., as Mosteller [11] has suggested). Ideally, we would know the true values of P(Xi > Xj) = pij, and, because the differences between the discriminal processes are Gaussian, we have

$$p_{ij} = \frac{1}{2\pi} \int_{-(S_i - S_j)}^{\infty} e^{-\frac{1}{2}y^2} dy$$

This can be solved (using the inverse error function) to find $(S_i - S_j)$, given ρ_{ij}. And, if we set S1 = 0, then we can compute all the other S_i.. This is the standard technique (we call it the Thurstone-Mosteller [TM] method), used in numerous studies [e.g., 3, 15].

In fact, however, we do not know the true values of $P(X_i > X_j)$. We have only estimates, based on the preference matrices for a small sample. The S_i thus have uncertainties associated with them (due to the stimuli), apart from the σ_i (due to the observer). Inferences about the set of S_i are difficult to obtain because of its unknown asymptotic distribution [1].

Bradley and Terry [13] and Bradley [9] formulated the problem somewhat differently (the BT method):

(1) t stimuli in an experiment using paired comparisons have true ratings.

$$p_i (i = 1, ..., t),\ p_i > 0$$

(2) Observations on pairs of treatments are independent in probability.

(3) When treatment i is compared with treatment j, the probability pij that treatment i is rated above treatment j is $p_i / (p_i + p_j)$.

Values of ln pi correspond to the values of S_i. Estimates of the p_i are found by maximum-likelihood methods, as described in [1, 9], using an iterative procedure. Moreover, confidence regions on the estimators of pi and ln pi are derived in [1, 9]. Bradley also defines a test statistic to allow testing of the null hypothesis that all of the p_i are equal against the alternative that some are unequal.

2 Methods

2.1 Data

Images for five outdoor scenes were acquired in the mid-wave and long-wave infrared (MWIR and LWIR). The imaging method yielded pairs of images that were registered spatially.

The five scenes are shown in Fig.1. In all cases, LWIR is on the left, and MWIR is on the right.

Image Fusion for Human Observers: How Should We Choose the Method?
Murray Loew, James Bonick, Clarence Walters
RDECOM, CERDEC, Night Vision and Electronic Sensors Directorate
Fort Belvoir, VA 22060

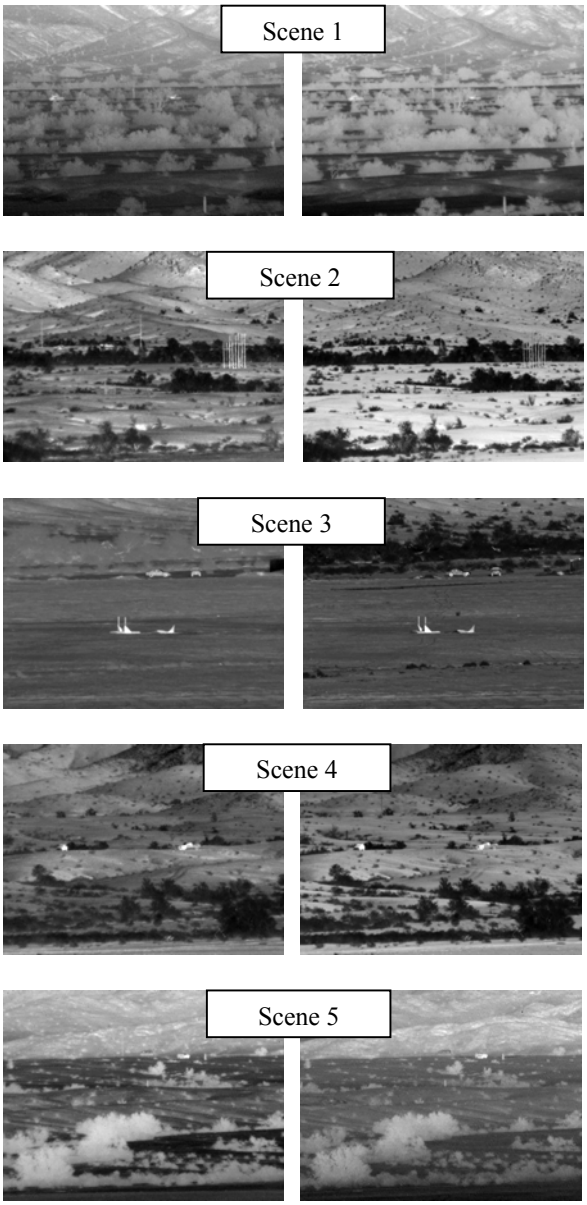


Figure 1. Scenes used in the study

2.2 Fusion Methods and Examples

Seven algorithms were used to fuse each pair of images; the sample pair below is used here to demonstrate the algorithms.


Figure 2. Sample images used to illustrate fusion methods

2.2.1 Image Averaging

The fused image (below) was a pixel-by-pixel average (equally weighted) of the MWIR and LWIR images. This was our baseline algorithm to assess any advantage provided by more-complex algorithms.

2.2.2 Maximum Pixel Value

Each pixel in the fused image below is the maximum of the MWIR and LWIR images' corresponding pixels. Would identification of these “hot spots” perform as well as more sophisticated techniques in distinguishing targets in fused images?

2.2.3 Wavelet

Wavelets are mathematical functions that are especially useful for describing signals at various degrees of resolution. The human visual system is sensitive to edges and corners. Multiresolution fusion methods take advantage of this. Since wavelet coefficients having large absolute values contain the information about the salient features of the images such as edges and lines, a good fusion rule is to take the maximum of the (absolute values of the) corresponding wavelet coefficients. [4, 5, 7]. We

computed the wavelet decomposition of both images (Daubechies 2 wavelet at 3 levels), combined them in the transform domain (using the maximum coefficients from each level) and reconstructed the fused image below [6].

2.2.4 Laplacian Pyramid

We computed the Laplacian pyramid decomposition of both images (3 levels) and reconstructed the fused image using the maximum coefficients from each level. Pyramid decomposition is another approach to representing images at multiple resolutions (scales). As noted above, because important features of images occur at various scales, it is useful to identify those features at each resolution, and make them evident in an image reconstructed from the several scales. To use this approach in fusion, we ask, at a given resolution, which of the images has the more-salient feature in a given region. This maximization is performed across the entire image, from which a combined image is created – at that resolution. The final image (below) is reconstructed from the set of combined images across the resolutions. A Laplacian pyramid uses that procedure, with sharpened images (Laplacian-filtered) as the building-blocks. As they are successively down-sampled (to reduce scale), an error measure is computed that is used in subsequent reconstruction of the image.

The rationale for this kind of fusion is that the resulting image will preserve and display the most-important parts of each image, at each scale.

2.2.5 Commercial False Coloring

This proprietary algorithm produced false coloring: the intensity in one band is mapped to the intensity of the fused image; intensity in the other band is mapped to color. (See image below.) The algorithm was implemented twice — first with LWIR providing intensity and MWIR providing color, and then reversed. The intent is to take advantage of human observers' rapid and reliable color perception.

2.2.6 Commercial Laplacian Pyramid

This algorithm is a proprietary variant of the Laplacian-pyramid fusion (also using three levels) described above. The algorithm also included proprietary noise-reduction and histogram-stretching operations; see image below.

2.3 Observer Studies

Each fusion method described above yielded one fused image for each of the five pairs of images. To the seven images resulting from each pair were added the pair itself. The data thus consisted, for each scene, of nine images (seven fused and two original). We wished to have human observers express a preference within each of all 36 possible pairs of the nine images for each of the five scenes – a total of 180 pairs. Image pairs were repeated (with position randomized) throughout the test, resulting in each participant making approximately 1000 judgments.

The test was performed by the University of Memphis at the request of the Modeling and Simulation Division of the U.S. Army Night Vision and Electronic Sensors Directorate. Six students participated; each had received training in target detection and identification prior to the test. For each pair of images, participants reported which image was of higher quality for the purpose of identifying the vehicles in the image.

The preference data were compiled into preference matrices; each matrix contained in each cell c(i,j) the number of times, for a given scene, that fusion method j was preferred to fusion method i, aggregated across all observers. In the example below, indices 8 and 9 refer to the original MWIR and LWIR images, respectively.

Table 1. A preference matrix: row indices denote Algorithm *i*; column indices denote Algorithm *j*. Each entry is the number of times that Algorithm *j* was preferred to Algorithm *i*.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	0	1	2	3	5	3	3	3
2	5	0	5	4	3	3	5	6	6
3	5	0	0	2	3	5	4	3	6
4	3	1	3	0	3	1	3	2	3
5	2	2	3	2	0	4	4	3	5
6	1	2	1	4	2	0	4	3	5
7	2	1	1	2	2	2	0	2	3
8	2	0	2	3	2	2	4	0	6
9	3	0	0	3	1	1	3	0	0

3 Results

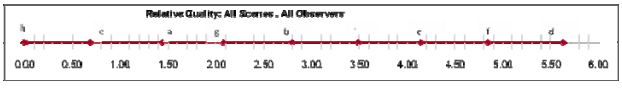
The preference-matrix data were analyzed using the TM and BT methods. Both return rankings of the fusion methods. Although the scale parameters are different for the two approaches, this does not affect the interpretation or comparison of the results. Because confidence intervals are not available for TM, the ranks are taken as computed. For BT, the ranks are recorded, confidence intervals computed, and error bars are applied so that a sense of the significance of the difference between ranks may be obtained.

3.1 TM Results

Each scene was analyzed separately, and then the preference matrices were combined to permit an assessment across all scenes. The results are tabulated below.

Table 2. TM results: Table displays preference measure for each algorithm for each scene, and over all scenes. Plot below illustrates the algorithm rank over all scenes, from least- to most-preferred (left to right).

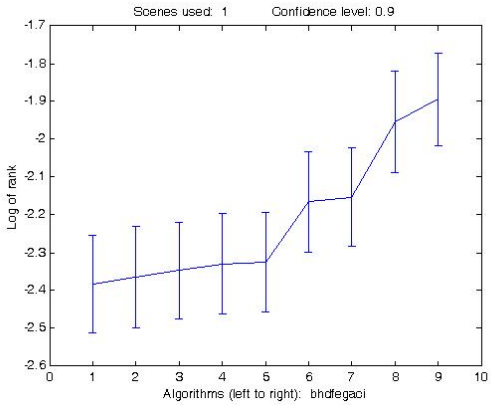
All Scenes		Scene 1		Scene 2		Scene 3		Scene 4		Scene 5	
Alg	Pref.	Pref.	Alg	Pref.	Alg	Pref.	Alg	Pref.	Alg	Pref.	Alg
h	0.00	0.00	b	0.00	a	0.00	h	0.00	a	0.00	i
e	0.85	0.69	d	0.48	e	0.70	g	0.85	e	0.57	g
a	1.58	1.44	h	0.98	h	1.40	e	1.60	h	1.26	c
g	2.25	2.07	e	1.74	b	2.12	f	2.34	c	2.10	e
b	2.95	2.79	f	2.53	c	3.01	c	3.10	i	2.72	a
i	3.86	3.49	a	3.37	i	3.77	a	3.72	b	3.51	h
c	4.56	4.14	g	3.92	g	4.57	i	4.67	f	4.16	f
f	5.23	4.84	c	4.70	f	5.34	d	5.54	g	4.79	b
d	5.91	5.63	i	5.44	d	5.90	b	6.17	d	5.35	d

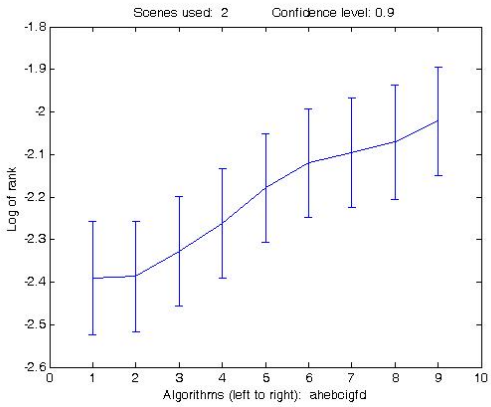


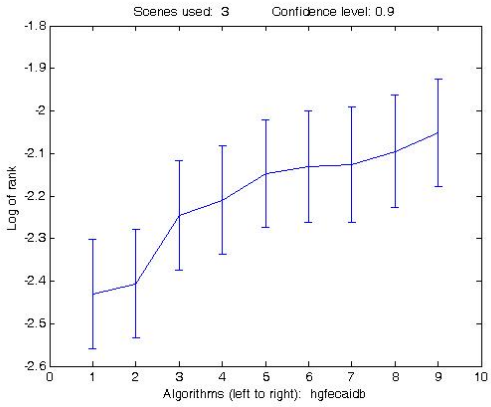
Algorithm key: a = Average; b = False color (MWIR intensity, LWIR color); c = False color (reversed); d = Laplacian pyramid; e = Maximum; f = Wavelet; g = Commercial Laplacian pyramid; h = original LWIR; i = original MWIR.

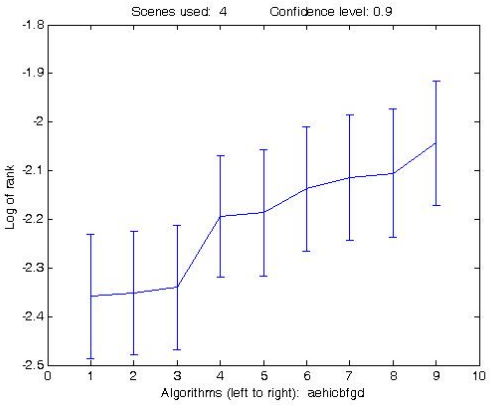
3.2 BT Results

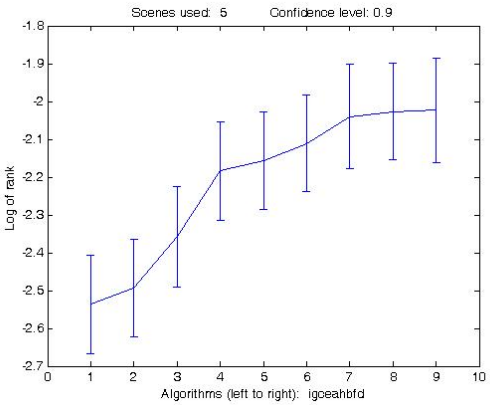
The figures below show, for each of the scenes, and then for all scenes taken together, the following: scene number; the algorithms sorted in increasing rank (left to right: note the algorithm code sequence below each plot); the logarithm of each algorithm's rank; and the width of the 90% confidence interval for each rank. It is thereby possible to identify groups of algorithms that are significantly different from other groups (see Sec. 3.3).

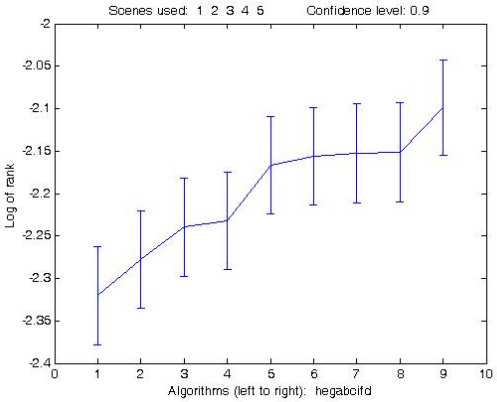












3.3 Comparison of Results

When the ranks reported by the two methods are compared, we observe great similarities. In the table below, many of the scenes have identical highest-rank sequences. The BT method, however, allows us to use the confidence-interval information to cluster algorithms that are not significantly different, and to identify clusters that are significantly different. It is evident, for example, that clusters involving algorithm d are preferred significantly more than those involving g, h, a, and e.

Table 3. Ranks computed by the two methods

	Low		RANK						High
Bradley-Terry									
Scene 1	b	h	d	f	e	g	a	c	i
2	a	h	e	b	c	i	g	f	d
3	h	g	f	e	c	a	i	d	b
4	a	e	h	i	c	b	f	g	d
5	i	g	c	e	a	h	b	f	d
12345	h	e	g	a	b	c	i	f	d
Thurstone									
Scene 1	b	d	h	e	f	a	g	c	i
2	a	e	h	b	c	i	g	f	d
3	h	g	e	f	c	a	i	d	b
4	a	e	h	c	i	b	f	g	d
5	i	g	c	e	a	h	f	b	d
12345	h	e	a	g	b	i	c	f	d

The two methods rank the algorithms similarly. Bradley-Terry also indicates which algorithms are clustered (blue, pink), and which clusters are significantly different (disjoint blue regions). Hatched areas are distinct, but not clustered.

Algorithm key: a = Average; b = False color (MWIR intensity, LWIR color); c = False color (reversed); d = Laplacian pyramid; e = Maximum; f = Wavelet; g = Commercial Laplacian pyramid; h = original LWIR; i = original MWIR.

4 Conclusions and Future Work

Overall, fusion does offer improvement over either infrared modality alone. The Laplacian pyramid was superior (except for Scene 1), and MWIR did well alone in Scenes 1, 2, and 3.

In practice, as has been noted in other studies [1, 2], TM and BT yield nearly identical scale estimates. If, however, a decision is to be made about which fusion algorithm to implement for a given application involving human preferences, one would like to know if there are any clear differences between algorithms. This is true not only because of the need to best serve human performance, but because if it were known reliably how much difference there was, it would then be possible to make a rational choice based on multiple factors (e.g., computation cost).

The limited sample size in the present case (six observers) led to relatively large confidence intervals. Increasing the sample size will shrink the confidence intervals, offering greater precision to the choice of algorithm. This measure of precision (at whatever level) is what gives the advantage to BT.
Another aspect of BT relates to the use of "complete" vs. "incomplete" data. The data used here were complete (i.e., all possible pairs were judged). As the number of algorithms and/or scenes grows, however, it becomes impracticable to present all possible pairs to the observers. There are several approaches to preference assessment that use subsets of the data, to which the BT model (but not TM) applies [1, 8].

Future work will involve experiments with fewer fusion methods and larger numbers of observers and scenes, and an attempt to understand what characteristics of Scene 1 set it apart with respect to observers' preferences.

5 Acknowledgments

We are grateful to the NVESD Modeling and Simulation Division and to the University of Memphis for their assistance in planning, and their conduct of, the experiments.

6 References

[1] J. C. Handley, “Comparative Analysis of Bradley-Terry and Thurstone-Mosteller Paired Comparison Models for Image Quality Assessment,” *Proc.* Image Processing, Image Quality, Image Capture Systems Conference (PICS-01), Montréal, Quebec, Canada, April 2001, [IS&T 2001], pp. 108-112.

[2] R. Bala, R. deQueiroz, R. Eschbach, and W. Wu, “Gamut Mapping to Preserve Spatial Luminance Variations,” *The Journal of Imaging Science and Technology*, Vol. 45, No. 5, Sept./Oct. 2001, pp. 436-443.

[3] J. Lanir, M. Maltz, S. Rotman, “Comparing multispectral image fusion methods for a target detection task,” *Optical Engineering*, Vol. 46, No. 6, June 2007, 066402-1 – 066402-8.

[4] P. Hill, N. Canagarajah, and D. Bull, "Image Fusion using Complex Wavelets," Electronic Proceedings of The 13th British Machine Vision Conference, University of Cardiff, 2-5 September 2002 (ed. by D. Marshall & P. L. Rosin). http://www.bmva.ac.uk/bmvc/2002/

[5] L.J. Chipman, T.M. Orr, and L.N. Lewis. Wavelets and image fusion. *IEEE Transactions on Image Processing*, 3:248–251, 1995.

[6] P.J. Burt and R.J. Kolczynski. Enhanced image capture through fusion. *Proceedings of the 4th International Conference on Computer Vision*, pages 173–182, 1993.

[7] Z. Zhang and R. Blum. A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. *Proceedings of the IEEE*, pages 1315–1328, August 1999.

[8] D. A. Silverstein and J. E. Farrell, "Efficient method for paired comparisons," *J. Electronic Imaging*, Vol. 10, No. 2, April 2001, pp. 394-398.

[9] R. A. Bradley, "Paired comparisons: some basic procedures and examples," Chap. 14 in *Handbook of Statistics, Vol. 4* (ed. by P. R. Krishnaiah and P. K. Sen), Elsevier Science Publishers, 1984.

[10] E. D. Montag, "Empirical formula for creating error bars for the method of paired comparison," *J. Electronic Imaging*, Vol. 15, No. 1, January-March, 2006, pp. 010502-1 – 010502-3.

[11] F. Mosteller, "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations," *Psychometrika*, Vol. 16, 1951, pp. 3-9.

[12] L. L. Thurstone, "Psychophysical analysis," *American J. Psychology*, Vol. 38, 1927, pp. 368-389.

[13] R. A. Bradley and M. E. Terry, "The rank analysis of incomplete block designs. I. The method of paired comparisons," *Biometrika*, Vol. 39, 1952, pp. 324-345.

[14] W. S. Torgerson, *Theory and Methods of Scaling*. New York: John Wiley & Sons, Inc., 1958.

[15] A. C. Copeland, M. M. Trivedi, and J. R. McManamey, "Evaluation of image metrics for target discrimination using psychophysical experiments," *Optical Engineering*, Vol. 35, No. 6, June 1996, pp. 1714-1722.