METHODS FOR IMAGE FUSION QUALITY ASSESSMENT – A REVIEW, COMPARISON AND ANALYSIS

Yun Zhang

Department of Geodesy and Geomatics Engineering
University of New Brunswick
Fredericton, New Brunswick, Canada
Email: YunZhang@UNB.ca;

Commission VII, WG VII/6

KEY WORDS: Remote Sensing, Digital, Comparison, Fusion, Accuracy

ABSTRACT:

This paper focuses on the evaluation and analysis of seven frequently used image fusion quality assessment methods to see whether, or not, they can provide convincing image quality or similarity measurements. The seven indexes are Mean Bias (MB), Variance Difference (VD), Standard Deviation Difference (SDD), Correlation Coefficient (CC), Spectral Angle Mapper (SAM), Relative Dimensionless Global Error (ERGAS), and Q4 Quality Index (Q4), which were also used in the IEEE GRSS 2006 Data Fusion Contest. Four testing images are generated to evaluate the indexes. Visual comparison and digital classification demonstrate that the four testing images have the same quality for remote sensing applications; however, the seven evaluation methods provide different measurements indicating that the four images have varying qualities. The image fusion quality evaluation by Alparone, et al. (2004) and that by the IEEE GRSS 2006 data fusion contest (Alparone, et al., 2007) are also analyzed. Significant discrepancy between the quantitative measurements, visual comparison and final ranking has been found in both evaluations. The inconsistency between the visual evaluations and quantitative analyses in the above three cases demonstrate that the seven quantitative indicators cannot provide reliable measurements for quality assessment of remote sensing images.

1. INTRODUCTION

Image fusion, especially the fusion between low resolution multispectral (MS) images and high resolution panchromatic (Pan) images, is important for a variety of remote sensing applications, because most remote sensing sensors, such as Landsat 7, SPOT, Ikonos, QuickBird, GeoEye-1, and WorldView-2, simultaneously collect low resolution MS and high resolution Pan images. To effectively fuse the MS and Pan images, numerous image fusion techniques have been developed with varying advantages and limitations. However, how to effectively evaluate image fusion quality to provide convincing evaluation results has been a challenging topic among the image fusion researchers and users of image fusion products.

In research publications, the widely used image fusion quality evaluation approaches can be included into two main categories:

(1) Qualitative approaches, which involve visual comparison of the colour between original MS and fused images, and the spatial detail between original Pan and fused images.
(2) Quantitative approaches, which involve a set of pre-defined quality indicators for measuring the spectral and spatial similarities between the fused image and the original MS and/or Pan images.

Because qualitative approaches—visual evaluations—may contain subjective factor and may be influenced by personal preference, quantitative approaches are often required to prove the correctness of the visual evaluation.

For quantitative evaluation, a variety of fusion quality assessment methods have been introduced by different authors. The quality indexes/indicators introduced include, for example, Standard Deviation (SD), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Sum Squared Error (SSE) based Index, Agreement Coefficient based on Sum Squared Error (SSE), Mean Square Error (MSE) and Root Mean Square Error, Information Entropy, Spatial Distortion Index, Mean Bias Error (MBE), Bias Index, Correlation Coefficient (CC), Warping Degree (WD), Spectral Distortion Index (SDI), Image Fusion Quality Index (IFQI), Spectral Angle Mapper (SAM), Relative Dimensionless Global Error (ERGAS), Q Quality Index (Q), and Q4 Quality Index (Q4) (e.g., Wald et al., 1997; Buntilov and Bretschneider, 2000; Li, 2000; Wang et al., 2002; Piella and Heijmans, 2003; Wang et al., 2004; Alparone et al., 2004; Willmott and Matsuura, 2005; Wang et al., 2005; and Ji and Gallo, 2006). However, it is also not easy for a quantitative method to provide convincing measurements. A commonly acceptable evaluation method has not yet been agreed by the authors of the quantitative evaluation papers.

In the practice of image fusion quality evaluation, it has been commonly noticed by researchers that the evaluation results can be affected (1) by the display conditions of the images when qualitative (visual) evaluation is conducted, and (2) by the selection of quantitative indicators (indexes) when quantitative assessment is performed.

- For visual evaluations, if a comparison is not conducted under the same visualization condition, i.e. if the images are not stretched and displayed under the same condition, the comparison will not provide reliable results. For example, an original MS image usually appears dark when no histogram stretching is applied, and it appears significantly differently when different stretches are applied (examples can be found in Figure 1). These different appearances are not caused by the quality difference, but just by the conditions of the image display. Therefore, one cannot conclude that one image is better than another if the display condition is not the same. Unfortunately, no display conditions were clearly described in many visual comparisons, including those in the IEEE GRSS 2006 Data Fusion Contest. This ambiguity in display conditions significantly reduced the reliability of the visual comparison results.

- For quantitative evaluation, different evaluation results can often be obtained when different quantitative measures or indicators are selected for the evaluation. Therefore, whether, or not, a given quantitative index can measure image fusion quality or measure quality difference between two images is still an open question. Among numerous quantitative evaluation indicators, the Mean Bias (MB), Variance Difference (VD), Standard Deviation Difference (SDD), Correlation Coefficient (CC), Spectral Angle Mapper (SAM), Relative Dimensionless Global Error (ERGAS), and Q4 Quality Index (Q4) have been often used in image fusion publications. They were also used in the IEEE GRSS 2006 Data Fusion Contest for quantitative evaluation.

Therefore, this paper focuses on the evaluation and discussion of how display conditions affect visual comparison and whether, or not, the seven often used quantitative indicators (MB, VD, SDD, CC, SAM, ERGAS, and Q4) can provide convincing results to tell the quality difference or similarity of two images. This evaluation is conducted based on the assumption that

(1) if two images of the same area can present identical information, including colour, spatial detail and image depth, under the same visualization condition, and

(2) if the two images can also provide the same classification result using the same classifier under the same processing condition,

the two images can be defined and accepted as having the same image quality.

This assumption is true for remote sensing imagery and remote sensing applications, because the two foremost important applications of remote sensing imagery are (1) visualization and (2) classification. If two images can provide the same results for visualization and classification under the same condition, they will not make any difference for remote sensing applications, and they can be equally accepted by remote sensing users.

For the evaluation and discussion, some testing images having the same image quality are generated; the seven quality indicators are applied to the testing images to check their ability to measure the quality similarity among the images; and the fusion quality evaluations by Alparone, et al. (2004) and Alparone, et al. (2007) are reviewed and analyzed to see whether, or not, the quality indicators of the evaluations provided convincing results.

2. TESTING IMAGES

An original Ikonos MS image of Fredericton, NB, Canada, collected on October 1, 2001, is used for the evaluation. The image contains 4 spectral bands and is stored in 16 bits. For visual comparison purpose and to test the performance of the seven quantitative indicators, the original Ikonos image (Ik-Orig) is altered through *mean shifting*, *histogram stretching*, and *histogram stretching plus mean shifting*, resulting in a mean shifted Ikonos image (Ik-Shift), a histogram stretched image (Ik-Str), and a histogram stretched and mean shifted image (Ik-Str-Shift). The detailed alteration of the Ikonos image is described in Table 1.

2.1 Visual comparison

To prove that the four images (Ik-Orig, Ik-Shift, Ik-Str, and Ik-Str-Shift) have the same image quality for visualization, they are displayed under the same display conditions and compared with each other. The histogram stretchings used are zero stretching (i.e. no stretching), linear stretching, root stretching, adaptive stretching, and equalization stretching (Figure 1). It can be seen that all of the four images appear very dark without any histogram stretching. And, all of the four images appear exactly the same when they are stretched using the same histogram stretch, regardless what stretch is applied (compare images in the same column of Figure 1). This comparison demonstrates that the four images have the same quality for visualization and visual interpretation.

On the other hand, it can also be seen from Figure 1 that the same image can be displayed and interpreted differently as if the source image had different qualities, if the image is not displayed under the same condition. For example, the same original Ikonos image (Ik-Orig) in Figure 1 appears significantly differently under different display conditions. Some appear darker than others, and some look noisier than others. If the image source information and the image stretching information were not given in Figure 1, one must say that the images in different columns of Figure 1 have different qualities.

Table 1. Alteration of the spectral bands of the original Ikonos MS (Ik-Orig) to obtain other testing images (Ik-Shift, Ik-Str, and Ik-Str-Shift)

	Ik-Orig	Ik-Shift	Ik-Str	Ik-Str-Shift
Band 1	B	B+100	B×1.5	B×1.5+100
Band 2	G	G+100	G×1.5	G×1.5+100
Band 3	R	R+100	R×1.5	R×1.5+100
Band 4	NIR	NIR+100	NIR×1.5	NIR×1.5+100

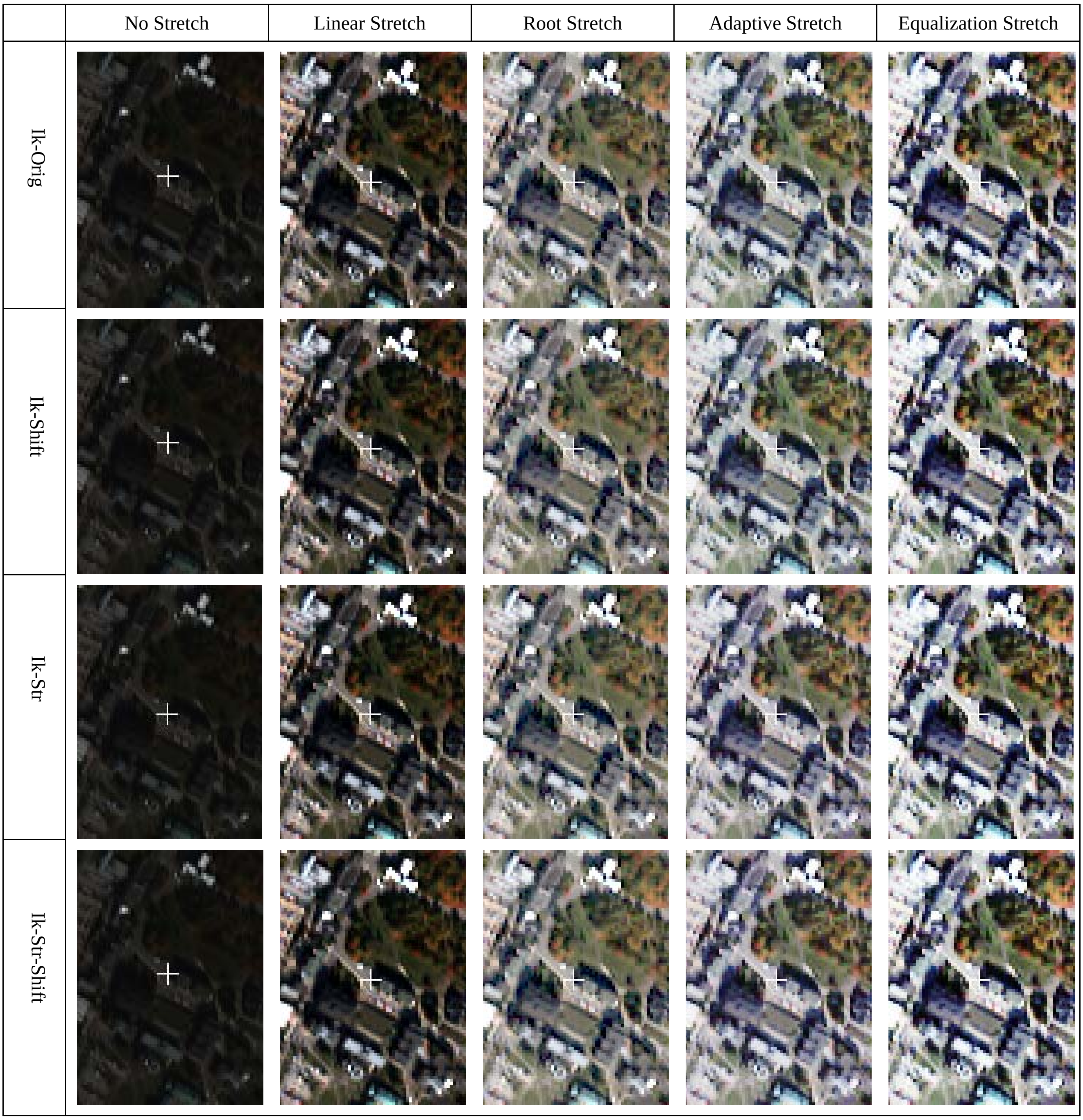


Figure 1. Comparison of the visual quality of the four testing images under the same display conditions (the images are enlarged 2 times to show details)

To prove that the four images have the same quality, the individual bands of each of the four testing images are also compared through overplaying the same band of different images in one image under the same stretching condition, but displayed using different colours (Figure 2). If there is any difference between corresponding bands of the images, the overlaid image will appear colour in the areas where differences exist. Otherwise, the overlaid bands will appear as a grey image, as if only one band were displayed. A close check of the images in Figure 2 shows that no colour appears in any of the images, which proves that the four images have the same quality when individual bands are compared.

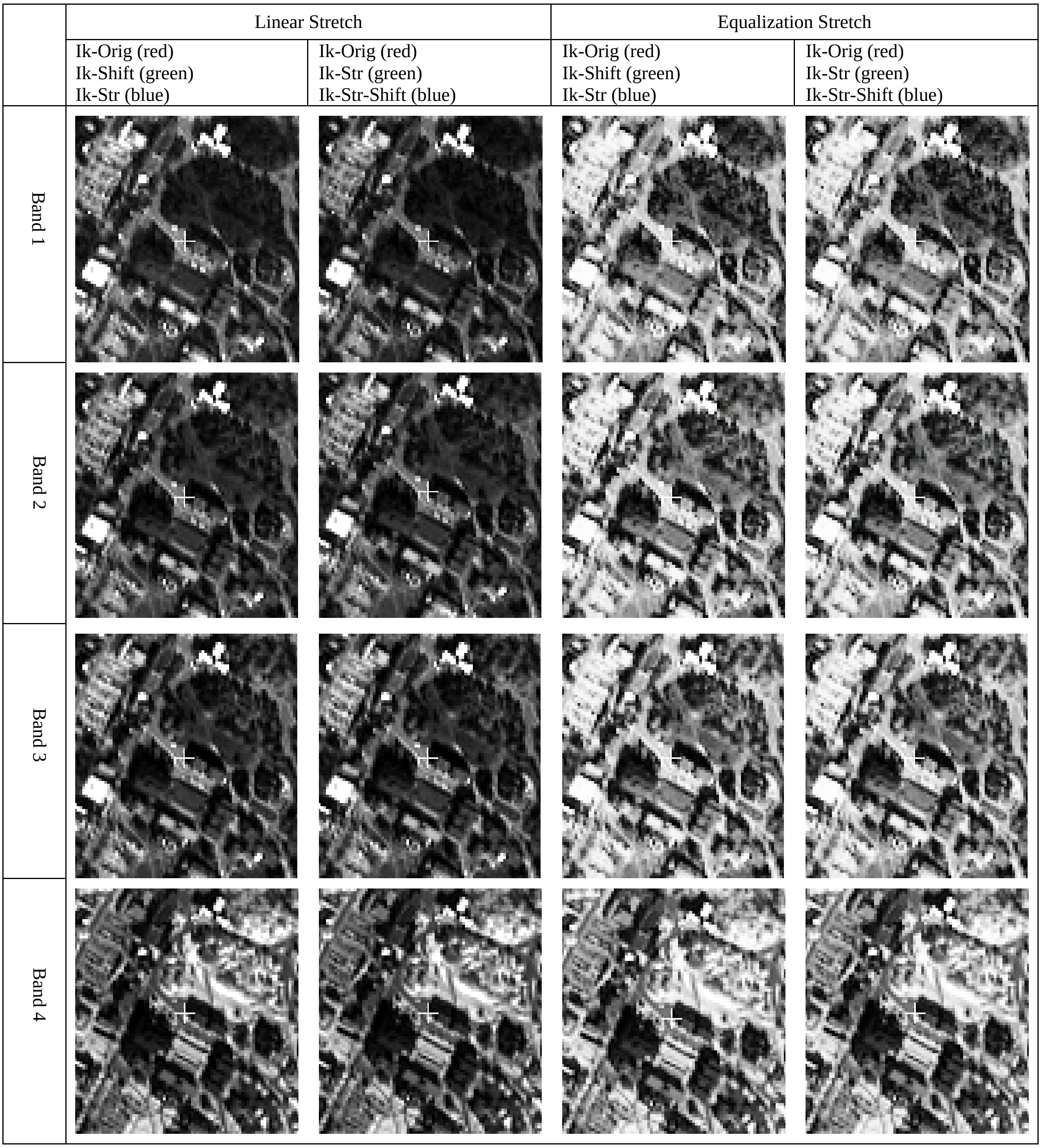


Figure 2. Comparison of quality difference between individual bands of the four testing images (Ik-Orig, Ik-Shift, Ik-Str, and Ik-Str-Shift) by overlaying the same band from different images in different colour and checking colour appearance in the overlaid images (colour indicating difference exists; no colour indicating no difference)

2.2 Classification

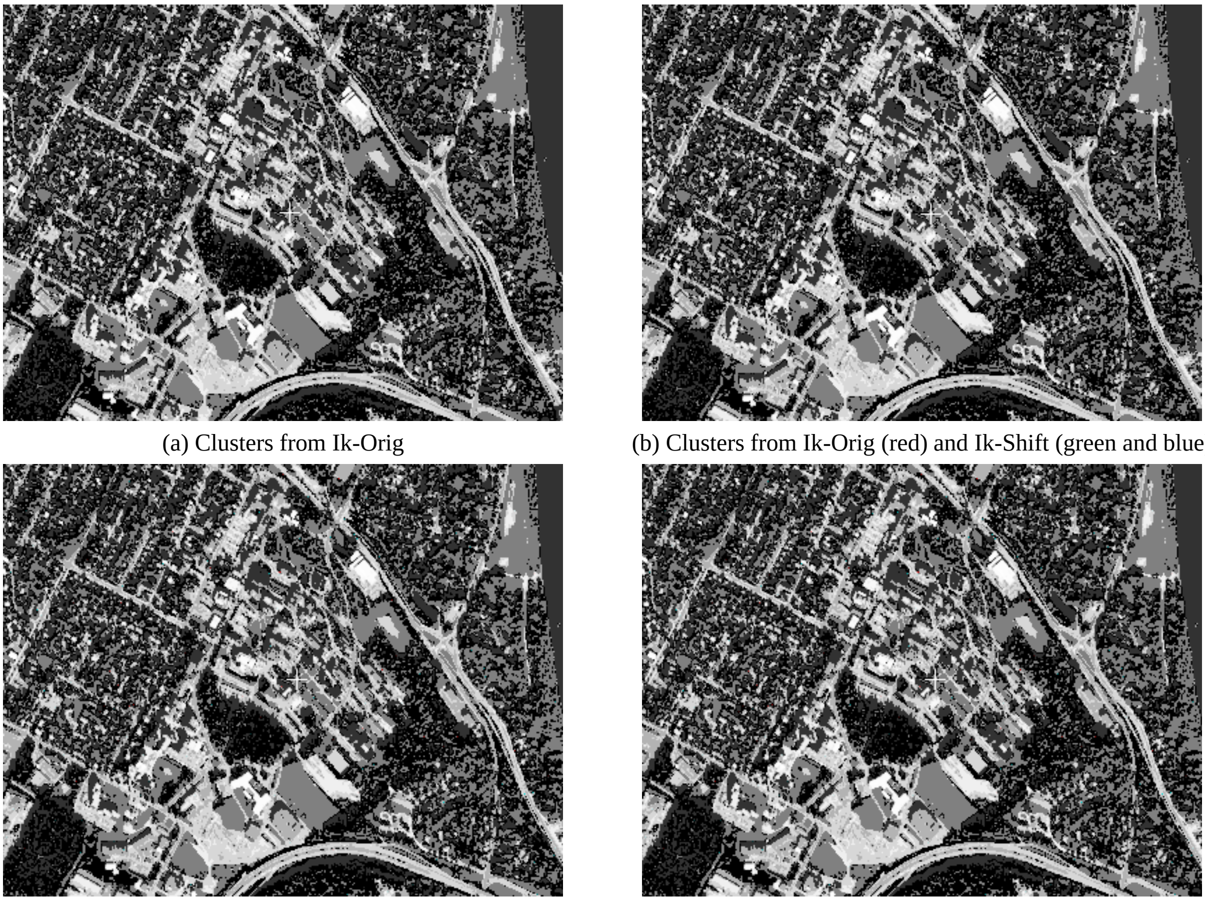
To prove that the four testing images have the same image quality for digital classification, the ISODATA clustering tool is selected to cluster the four images into the same number of clusters using exactly the same clustering parameters. The ISODATA clustering is selected, instead of any supervised classifiers, to avoid operator’s influence in the classification. All of the four spectral bands are used in the clustering. The images are clustered into 16 clusters. And, the maximum clustering iteration is 20.

To precisely compare the 16 clusters classified from the four images, the clustering result from Ik-Orig is shown in Figure 3.a, and the results from Ik-Shift, Ik-Str and Ik-Str-Shift are overlaid with that of Ik-Orig and displayed in Figure 3.b, 3.c and 3.d, respectively. It can be seen that all of the clustering results appear the same as that of the Ik-Orig. No colour appears anywhere in the overlaid clustering results (Figure 3.b, 3.c and 3.d). This comparison proves that the images Ik-Orig, Ik-Shift, Ik-Str and Ik-Str-Shift have the same quality for image classification.

The check of the statistic reports of the clustering results shows that the pixel number in each of the 16 clusters are exactly the same in the results of Ik-Orig and Ik-Shift. A few outlier pixels are identified when compare between the statistic results of Ik-Orig and Ik-Str. Out of the 16 clusters, 9 clusters are identical, and 7 clusters have a few pixels of difference. In total, 98 pixels are identified as outliers in the classification/clustering of an image with more than 1 million of pixels (1024×1024 pixels). The outliers may be caused by limited clustering iterations and/or other settings. Therefore, the statistic reports also demonstrate that the four testing images have the same quality.

3. EVALUATION OF THE SEVEN QUANTITATIVE INDICATORS

To evaluate the capability of the seven often used quantitative indicators (MB, VD, SDD, CC, SAM(°), ERGAS, and Q4) for the measurement of image similarity or difference, they are applied to the three testing images Ik-Shift, Ik-Str and Ik-Str-Shift with the image Ik-Orig as reference. The quality measurement values are shown in Table 2.



(a) Clusters from Ik-Orig

(b) Clusters from Ik-Orig (red) and Ik-Shift (green and blue)

(c) Clusters from Ik-Orig (red) and Ik-Str (green and blue)

(d) Clusters from Ik-Orig (red) and Ik-Str-Shift (green and blue)

Figure 3. Comparison of the 16 clusters clustered from the four testing images using ISODATA method (in (b), (c) and (d) the clusters from Ik-Shift, Ik-Str and Ik-Str-Shift are displayed in green and blue and overlaid with that of Ik-Orig displayed in red)

Table 2. Values of the seven image quality indicators for similarity measurements between the reference image Ik-Orig and the three testing images Ik-Shift, Ik-Str and Ik-Str-Shift

	MB	VD	SDD	CC	SAM(°)	ERGAS	Q4
Ik-Shift	100	0	0	1	9.047	6.159	0.950
Ik-Str	129.076	10582.9	41.398	1	9.725	8.584	0.852
Ik-Str-Shift	229.076	10582.9	41.398	1	15.918	6.793	0.766

MB – Mean Bias, VD – Variance Difference, SDD – Standard Deviation Difference, CC – Correlation Coefficient, SAM – Spectral Angle Mapper, ERGAS – Relative Dimensionless Global Error; Q4 – Q4 Quality Index
(Ideal values: MB = 0; VD = 0; SDD = 0; CC = 1; SAM(°) = 0; ERGAS = 0; Q4 = 1)

According to Alparone et al. (2007), if there is no quality difference between two images, the value of Mean Bias (MB), Variance Difference (VD), Standard Deviation Difference (SDD), Spectral Angle Mapper (SAM), and Relative Dimensionless Global Error (ERGAS) should be zero, and the value of Correlation Coefficient (CC) and Q4 Quality Index (Q4) should be one. Larger values of MB, VD, SDD, SAM, and ERGAS indicate larger quality difference between two images. For CC and Q4, however, the worst value is zero.

According to the evaluation criteria of Alparone et al. (2007) and comparing the values in Table 2, we can find that:

- Four out of the seven indexes (MB, SAM, ERGAS and Q4) indicate that the three images Ik-Shift, Ik-Str and Ik-Str-Shift have different quality than that of Ik-Orig.
- Two others (VD and SDD) indicate that Ik-Shift has the same quality as Ik-Orig, whereas Ik-Str and Ik-Str-Shift have different quality than Ik-Orig.
- Only one out of the seven indexes (CC) indicates that all of the four images Ik-Orig, Ik-Shift, Ik-Str, and Ik-Str-Shift have the same image quality.

With such a significant disagreement between the seven indexes, can they still measure the quality difference or similarity of two images? If yes, which index should we rely on and how can we explain the disagreement?

On the other hand, if the seven indexes could tell the quality difference between two images, i.e. a fused image and the original MS image, one should be able to easily improve the values of the measurements by just systematically shifting the means of the fused images to the desired means of the original MS images, and/or by systematically stretching the histograms of the fused images to match the desired standard deviation of the original MS images. Do these systematic adjustments and the improvements of the measurement values actually improve the quality of the image fusion results? Definitely not.

4. DISCREPANCY OF SAM, ERGAS, Q4 AND CC EVALUATION

Alparone et al. (2004) introduced a global quality measurement —Q4 Quality Index (Q4)—for image fusion quality evaluation, because the ERGAS method failed in measuring spectral distortion.

In the evaluation of Alparone, et al. (2004), QuickBird MS and Pan images were first degraded from 2.8m and 0.7m to 11.2m and 2.8m respectively. The degraded MS and Pan images were then fused to obtain pan-sharpened 2.8m MS images. The original 2.8m MS image was used as a reference image (or ground truth) to compare with the pan-sharpened MS images for quantitative measurement of the fusion quality. The image fusion methods evaluated were HPF (High Pass Filter), IHS, GLP-SDM (Alparone et al., 2003) and GLP-CBD (Alparone et al., 2003) methods. In addition, the degraded 11.2m MS image (denoted as EXP) and a modified 2.8m MS image (denoted as SYN) were also compared with the original 2.8m MS image for quantitative measurements of the image quality. The modified 2.8m MS image (SYN) was generated by multiplying the 4 spectral bands of the original 2.8m MS image with a constant 1.1. The quantitative measurements are cited in Table 3.

According to the measurement values in Table 3, we can see that SYN results should be the best (better than the GLP-SDM and GLP-CBD results), because:

- SYN has the highest CC value, 1;
- SYN has the highest Q4 value, 0.991 (closest to 1);
- SYN has the smallest SAM value, 0°, no spectral distortion was introduced; and
- although SYN has a higher ERGAS value than GLP-SDM and GLP-CBD do, this value should not be overly concerned, because ERGAS failed in measuring spectral distortion according to Alparone et al. (2004).

When readers compare the SYN, GLP-SDM and GLP-CBD images with the reference image (original 2.8m MS image) displayed in Alparone et al. (2004), readers can also see that the SYN results have the best quality, because the SYN image is closest to the original true 2.8m MS image in terms of spectral and spatial information, whereas the GLP-SDM image contains significant colour distortion and GLP-CBD image is blurred.

However, Alparone et al. (2004) stated in the final ranking that the results of SYN were confusing if ERGAS was compared to Q4, and both the GLP-SDM and the more sophisticated GLP-CBD results were the best according to the Q4 index and correlation measurements. How can readers understand this ranking? Was this ranking a result of the quantitative measurements, the visual comparison, or personal preference?

Table 3. Quality measurements of the pan-sharpened images (HPF, IHS, GLP-SDM, and GLP-CBD), low resolution MS image (EXP) and modified MS image (SYN) with the original MS image as reference (data source: Alparone et al. (2004))

	EXP	**SYN**	HPF	IHS	GLP-SDM	GLP-CBD
CC_{Ave}*	0.845	**1**	0.814	0.717	0.823	0.912
Q4	0.756	**0.991**	0.876	0.864	0.885	0.909
SAM(°)	2.17	**0.00**	2.54	2.97	2.17	1.64
ERGAS	1.793	**2.292**	1.943	2.540	1.579	1.180
* CC_{Ave} = average CC of the four spectral bands (calculated according to Table III of Alparone et al. (2004))						

5. PROBLEMS IN THE IEEE GRSS 2006 DATA FUSION CONTEST

5.1 Background

Giga bites of testing data were made available to the contest participants for image fusion. The testing data contain two types of images: (1) QuickBird and (2) simulated Pleiades Pan and MS images. The Pleiades Pan images were simulated using green and red channels, which did not cover the designed spectral coverage of Pleiades Pan, 500-850 nm, analogously to Ikonos and QuickBird Pan (Alparone et al. 2007). The data volume of QuickBird images occupies over 80% of the total data volume provided to the participants for the contest. The fusion results generated by the participants were sent to one of two official contest judges, Dr. L. Alparone.

5.2 Contest results

The contest evaluation concluded that the fusion results of the *Generalized Laplacian Pyramid Decomposition Featuring a Modulation Transfer Function Reduction Filter and a Context Based Decision Injection Rule* (GLP-MTF-CBD), also called GLP-CBD, outperformed the other competing algorithms for most of the criteria [MB, VD, SDD, CC, SAM, ERGAS, and Q4] (Gamba et al., 2006). An IEEE Certificate of Recognition was granted to the GLP-CBD developers at the IEEE IGARSS 2006 conference in August 2006.

The paper on 2006 data fusion contest outcome (Alparone et al., 2007), published in *IEEE Transactions of Geoscience and Remote Sensing*, provided results of quantitative analysis and visual evaluation. The visual analysis stated:

- *"GLP-CBD:* Image is nice as a whole. Colors should be better synthesized. This would enhance the legibility of the image. Details are there, except for the most colored (blue, red). Errors in colors lead to interpretation errors. Contours should be sharper. There is no bias, except for Strasbourg outskirts. Unacceptable for detailed visual analysis."
- *"UNB-Pansharp:* Image is too noisy. There are many artifacts. Colors are not well synthesized as a whole and locally. Green trees are not green enough. Red or blue cars are absent. Shapes are not well defined; they are sometimes underlined by black lines. Too large bias is observed. There is lack of variance as a whole. At times, unacceptable. In best cases, unacceptable for detailed visual analysis."

5.3 Irregularity of the evaluation

After the contest award in August 2006, numerous requests were sent to the contest committee for an opportunity to review some fusion examples by the contest participants. The requests were rejected and the participants were asked to wait for the publication of the paper on the contest outcome. Finally, the evaluation examples were provided to the participants for review in late December 2006.

In the reviewing of the fusion results used in the contest evaluation, it was found that the QuickBird fusion results produced by UNB-Pansharp were not evaluated in the contest, even though giga bytes of QuickBird fusion results of UNB-Pansharp were sent to the judge, Dr. L. Alparone, together with fusion results of the simulated Pleiades data.

Two subsets of the UNB QuickBird fusion results are given in Figure 4. Readers can compare the original QuickBird Pan and MS images with the fusion results to evaluate whether the visual analysis of the IEEE fusion contest outcome by Alparone, et al. 2007 (see above) is objective, or not. Internal evaluation among the contest participants clearly agreed that the results produced by UNB-Pansharp are superior to those of GLP-CBD.

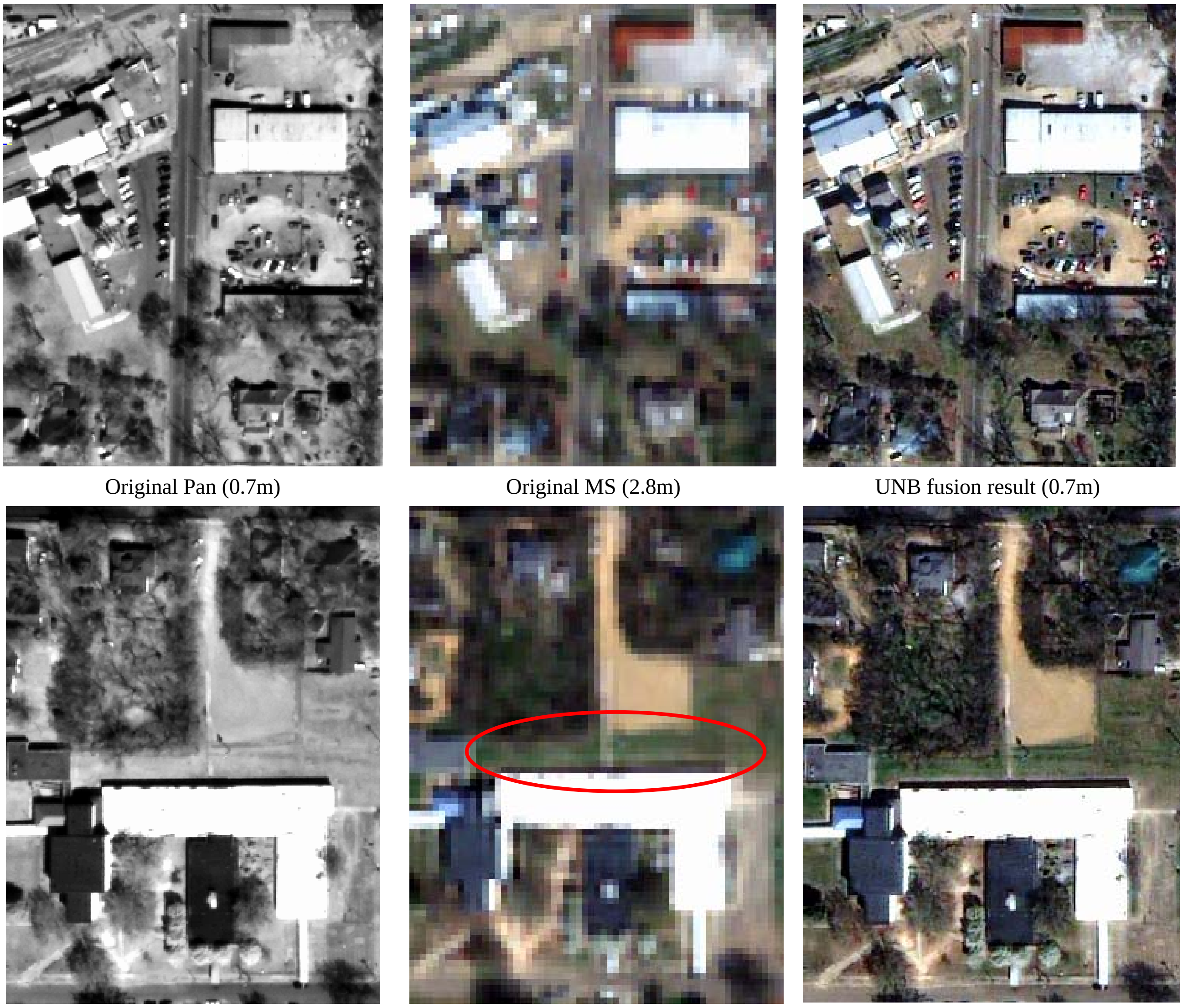
Literature review after the fusion contest, especially after the publication of the contest outcome (Alparone et al., 2007), proves that the judge Dr. L. Alparone is also a co-developer/co-author of the top winning GLP-CBD algorithm (Alparone et al., 2003; Aiazzi & Alparone et al., 2002; and Aiazzi & Alparone et al., 2006).

A request for permission to use the GLP-CBD fusion results provided to the 2006 contest participants for publications was denied. The original answer to the participants is quoted below to avoid misinterpretation:

> "In particular, on the Quickbird fused image I noted few small areas where a proper spatial enhancement did not occur because of statistical instabilities of the adaptive injection model. Such fusion inaccuracies appear only in few small areas and cannot change the global evaluation of my algorithm. However, il [if] one extracts the misfused patches and compares only them with those of other algorithms, he might be erroneously lead to believe that GLP-CBD is not the best algorithm among those compared in the Contest. After the contest I realized of the inconvenience by watching the fused images in the DFC [data fusion contest] web site and I fixed it. On the other side, the DFC site should contain the images that were evaluated for the Contest and cannot be changed. Therefore, if you want use GLP-CBD fused data for any publications, I will be pleased to provide fused versions with the fixed algorithm, which performs identically to the earlier one except on the above mentioned small areas.
>
> So, I do not give you, or any other may request it, the permission of using the GLP-CBD fused data found in the DFC contest, because such data refer to the Contest only and do not reflect the current progress of my activity, as it should appear in an unbiased future publication."

In the comparison between the GLP-CBD QuickBird fusion results received by the contest participants in 2006 and that published in the IEEE contest outcome paper (Alparone et al., 2007), it is found that the GLP-CBD QuickBird fusion result in Alparone et al. (2007) is clearly better than the one received by the participants—misfused patches and blurred areas are clearly reduced.

Because the request for permission was rejected, the comparison between the GLP-CBD QuickBird fusion result used in the contest in 2006 and that published in the contest outcome paper in 2007 (Alparone et al., 2007) cannot be displayed here. However, readers can still see some difference by comparing the GLP-CBD QuickBird fusion result published in the IEEE GRSS Newsletter (Gamba et al., 2006) and that in the contest outcome paper (Alparone et al., 2007), even though the images displayed are very small and do not cover the same area. To see the misfused patches, readers can see the GLP-CBD QuickBird fusion result published in the IEEE GRSS Newsletter (Gamba et al., 2006) and pay attention to the area circled in Figure 4 of this paper. The difference leads to a question: how can the GLP-CBD QuickBird fusion result published in the contest outcome paper in 2007 appear significantly better than that submitted to the contest in 2006?



Original Pan (0.7m)

Original MS (2.8m)

UNB fusion result (0.7m)

Original Pan (0.7m)

Original MS (2.8m)

UNB fusion result (0.7m)

Figure 4. Subsets of the QuickBird fusion results of UNB-Pansharp submitted to the IEEE GRSS 2006 Data Fusion Contest (UNB-Pansharp can produce fusion results either with or without feature enhancement. The fusion results with feature enhancement were submitted to the contest. All images in this figure are displayed under the same image stretching condition.)

The inconsistency and irregularity in the evaluation of IEEE GRSS 2006 Data Fusion Contest also raised the question on the capacity of the seven quantitative indicators (MB, VD, SDD, CC, SAM, ERGAS, and Q4) for quality measurements between images.

6. CONCLUSIONS

This paper analyzed and evaluated three cases of image quality comparisons using visual and quantitative methods. The three cases are (1) visual and quantitative analysis of the four testing images generated for this study; (2) review and analysis of the fusion quality evaluation by Alparone et al. (2004), which received the 2004 *IEEE Geoscience and Remote Sensing Letter* Best Paper Award (Alparone et al., 2007); and (3) review and analysis of the evaluation of the IEEE GRSS 2006 data fusion contest. The quantitative methods evaluated are the seven frequently used indicators—Mean Bias (MB), Variance Difference (VD), Standard Deviation Difference (SDD), Correlation Coefficient (CC), Spectral Angle Mapper (SAM), Relative Dimensionless Global Error (ERGAS); Q4 Quality Index (Q4)—which are also the quantitative measures of the IEEE GRSS 2006 Data Fusion Contest.

In the visual and quantitative analysis of the four testing images generated for this study, it was found:

- The four testing images generated through mean shifting and/or histogram stretching provide the same visualization and classification results under the same display and classification conditions. This demonstrates that mean shifting and histogram stretching (within the allowed digital number range of the file) do not change image quality for remote sensing applications.

- Visual evaluation results can be strongly influenced by image display conditions. The same image can be interpreted as having different qualities, if the display conditions are not the same. Therefore, it is important to assure a consistent display condition for images compared to achieve a convincing visual comparison result.
- Significant disagreement exists in the quantitative measurements of the seven indicators. Images having the same quality for remote sensing applications are indicated as having significant quality difference. This proves that the indicators are not capable of providing convincing image similarity measurements.

In the image fusion quality evaluation by Alparone et al. (2004), the SYN result is clearly the best according to the Q4, SAM and CC measurements, as well as the visual comparison. Although the SYN result does not have the best ERGAS value, it should not be overly concerned because according to Alparone et al. (2004) ERGAS failed in measuring spectral distortion. However, in the final ranking of Alparone et al. (2004), the authors' fusion algorithms GLP-SDM and GLP-CBD were ranked as the best, instead of the SYN results. This demonstrated that the authors themselves did not trust the measurement values, and personal preference played an important role in the ranking.

In the fusion quality evaluation of the IEEE GRSS 2006 Data Fusion Contest, the inconsistency and irregularity of the evaluation has suggested the difficulty of using the seven quantitative indicators to provide convincing quality measurements. Otherwise, there would have been no need to be selective in the contest evaluation for showing that the judge's GLP-CBD algorithm was the best and first class in the fusion contest, and the obvious, misfused patches or areas would have been detected.

In conclusion, the discrepancy between the visual evaluations and quantitative analyses in the three cases discussed in this paper demonstrate that the seven quantitative indicators (MB, VD, SDD, CC, SAM, ERGAS, and Q4) cannot provide reliable measurements for quality or similarity assessment between remote sensing images.

ACKNOWLEDGEMENTS

The author thanks Mr. Z. Xiong and Mr. J. D. Mtamakaya for their kind support in data and material preparation. The author also thanks the City of Fredericton, NB, Canada, for providing the original Ikonos Pan and MS images, and the IEEE GRSS 2006 data fusion contest committee for the original QuickBird Pan and MS images.

REFERENCES

Aiazzi, B., L. Alparone, S. Baronti, A. Garzelli, and M. Selva, 2006. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogrammetric Engineering and Remote Sensing*, Vol. 72, No. 5, pp. 591–596.

Aiazzi, B., L. Alparone, S. Baronti, and A. Garzelli,, 2002. Context-driven fusion of high spatial and spectral resolution data based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, No. 10, pp. 2300–2312.

Alparone, L., B. Aiazzi, S. Baronti, A. Garzelli, 2003. Sharpening of very high resolution images with spectral distortion minimization. *Proceedings of 2003 IEEE International Geoscience and Remote Sensing Symposium* (IGARSS 2003), pp. 458- 460.

Alparone, L., B. Aiazzi, S. Baronti, A. Garzelli, and P. Nencini, 2004. A Global Quality Measurement of Pan-Sharpened Multispectral Imagery. *IEEE Geoscience and Remote Sensing Letters*, Vol. 1, No. 4, October 2004. pp. 313-317.

Alparone, L., L.Wald, J. Chanussot, C. Thomas, P. Gamba, L.M. Bruce, 2007. Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data-Fusion Contest. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 45, No. 10, Oct. 2007, pp. 3012 – 3021.

Buntilov, V. and T. Bretschneider, 2000. Objective Content-Dependent Quality Measures for Image Fusion of Optical Data. *International Archives of Photogrammetry and Remote Sensing*, Vol. 33, 2000.

Gamba, P., J. Chanussot, and L. M. Bruce, 2006. TECHNICAL COMMITTEE REPORTS: Contest Organized by the Data Fusion Technical Committee at IGARSS 2006. *IEEE Geoscience and Remote Sensing Society Newsletter*, December 2006, pp. 11 – 16.

Ji, L., and K. Gallo, 2006. An Agreement Coefficient for Image Comaparison. *Photogrammetric Engineering and Remote Sensing Journal*, Vol. 72, No. 7, pp. 823-833.

Li, J. 2000. Spatial Quality Evaluation of Fusion of Different Resolution Images. *International Archives of Photogrammetry and Remote Sensing*, Vol. 33, 2000.Piella, G., and H. Heijmans, 2003. A new quality metric for image fusion. *Proceedings of IEEE International Conference on Image Processing*, Vol. 3, pp.173-176.

Wald, L., T. Ranchin, and M. Mangolini, 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, Vol. 63, No. 6, pp. 691–699.

Wang, Z, D. Ziou, C. Armenakis, D. Li, and Q. Li, 2005. A Comparative Analysis of Image Fusion Methods. *IEEE Transactions on Geoscenc and Remote Sensing*, Vol. 43, No. 6, pp.1391-1402.

Wang, Z., A.C. Bovik, H. Sheik, and E. Simoncelli, 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600-612.

Wang, Z., and A.C. Bovik, 2002. A Universal Image Quality Index. *IEEE Signal Processing Letters*, Vol. 9, No.3, pp.81-84.

Willmott, C. and K. Matsuura, 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing the average model performance. *Climate Research*, Vol. 30. pp. 79-82.