Statistical Analysis of the Performance Assessment Results for Pixel-Level Image Fusion

Zheng Liu
Intelligent Information Processing Laboratory
Toyota Technological Institute
Nagoya, Aichi, 468-8511 Japan
Email: zheng.liu@ieee.org

Erik Blasch
Information Directorate
Air Force Research Laboratory
Rome, NY, 13441 USA
Email:erik.blasch.1@us.af.mil

***Abstract*—Pixel-level image fusion (PLIF) performance assessment includes information theory, feature-based, structural similarity, and perception-based objective metrics. However, to relate these metrics to human understanding requires subjective metrics. This paper proposes to use statistical analyses to assess PLIF performance over objective and subjective metrics. Non-parametric tests are applied to the subjective and objective assessment data from three multi-resolution image fusion methods using visual and infrared images. The tests can offer the performance information about the fusion algorithm at a designated significance level. Statistical analysis of PLIF facilitates the establishment of a baseline for the research in image fusion and serves as a statistical validation for proposing, comparing, and adopting a new PLIF algorithm.**

I. Introduction

Pixel-level fusion of multi-sensor images has found a diverse range of applications, including surveillance, driving assistance, medical diagnosis, industrial inspection etc. [1]. Many image fusion algorithms have been proposed and the performance of these algorithms needs to be verified, assessed, validated, and compared. Currently, the assessment is conducted by using image fusion performance metrics over information theory, image feature-based, structural similarity, or human perception-based objective measures. These metrics are based on a computational model counting the amount of image feature, content, or information transferred from inputs to the fused result [2]. A comprehensive assessment will tell which fusion algorithm performs better for a specific set of sensors, targets, and environment data or applications [3]. However, the accuracy and reliability of the fusion metric is not complete [4]. Each metric may reveal one inherent property of the fusion process or the fused image. Sometimes, multiple metrics may give a contrary judgments. Thus, it is necessary to set up a baseline evaluation that brings together objective measures, common data sets, and subjective human reasoning for a consistent comparison of methods. Human subjective assessment plays a paramount role in the fusion performance assessment [5].

In pilot work done by European researchers, subjective assessment data was collected [6]. Within a precision-recall framework, the fusion results were assessed based on the comparison with its inputs, with which a reference was generated. The comparison is implemented with the semantic segmentation instead of the original images. This work proposed an idea on how to establish a ground truth reference to assess fused results. However, how to use the subjective assessment data has not been fully exploited. A closely related issue is about the fusion metric itself, which can be of quantitative or qualitative representations [7]. Once such a ground truth is set up, it is possible to learn how successful a fusion metric performs in reporting on the image fusion performance.

Statistical analysis has been used to compare different algorithms over multiple data sets in the research of machine learning [8], [9], [10], [11], [12]. Different classifiers can be compared with a set of non-parametric statistical tests. With the tests, it is possible to decide which classification algorithm is considered better than another. When a new algorithm is proposed, it is good to have a baseline method from which to compare the current performance against other techniques in a similar scenario.

We seek to employ statistical methodologies for the image fusion performance assessment. More specifically, the statistical analysis should be able to tell which fusion algorithm performs better under specific conditions. We use the data sets presented in [6] to discuss subjective and objective assessment. The performance of selected fusion algorithms were analysed with the non-parametric statistical tests on subjective assessment and objective fusion metrics. The test results lead to some discussions about the subjective assessment and the use of fusion metrics, which may serve as a reference for future research and development work.

This paper is organized as follows. Section II discusses PLIF performance assessment and Section III overviews the statistical tests. Section IV details experimental results. Section V provides a discussion and Section VI draws conclusions.

II. Pixel-Level Image Fusion and Performance Assessment

A. Pixel-level image fusion (PLIF)

The purpose of pixel-level image fusion (PLIF) is to create a composite image, which incorporates the most salient information or features from the input images [1]. The end user or application can benefit from the use of this fused image for perceivability, image enhancement, and target recognition. Pixel-level fusion has found its applications in a diverse fields, such as surveillance, photography, industrial inspection, and medical diagnosis etc. There are three categories of sensor fusion. Signal-level fusion uses the data content (e.g., pixels); whereas feature-level and decision-level fusion utilize

exploited attributes (e.g., intensity) and decision scores (e.g., probabilities). The limitation of PLIF is the computational cost to fuse all data points and the benefit of PLIF is that a product such as a fused image is available for user inspection.

There are numerous algorithms have been proposed to implement PLIF, among which the multi-resolution analysis (MRA) based approaches have demonstrated great potential and performance. The basic procedure is illustrated in Fig. 1. The image available to the MRA methods are spatially registered images which can be over different modality collections such as visual and thermal bands. The MRA algorithms perform an image decomposition using a variety of structured image pyramids and wavelet transforms. The key technique is the coefficient combination in the transform domain which is a method of image fusion. Using these coefficients, the image has to be reconstructed using the inverse of the image pyramid or wavelet transform chosen for decomposition. Finally, a rendered fused image is available to the user and the optimized image is subject to user, application, ad data-availability needs. The detailed information in [1], [13], [14], [15] provide good references for the reader interested in more details of the MRA methods.

The data presented in [6] was used in this study, which is also available from the image fusion website [16]. In this data set, three fusion algorithms were implemented, including a pyramid-based (PYR) approach, a discrete wavelet transform (DWT), and a complex wavelet transform (CWT) [15]. An example from this data set is given in Fig. 2. The input visible image and infrared image were fused by the aforementioned three algorithms.

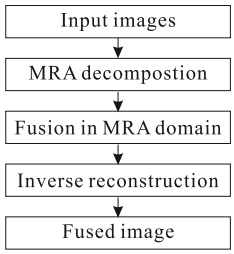
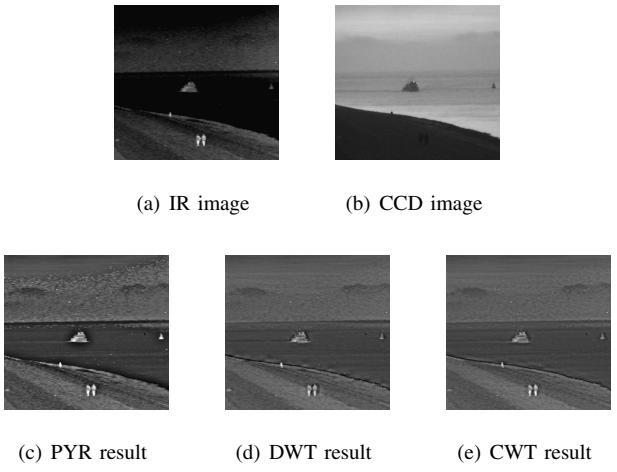


Fig. 1. MRA-based pixel-level fusion.



(a) IR image (b) CCD image

(c) PYR result (d) DWT result (e) CWT result

Fig. 2. An example of pixel-level image fusion.

Given these images, it is important to compare and contrast the various methods to determine which method is the best to use over various conditions, data quality, user preferences, and target discernability.

B. Fusion performance assessment

Performance assessment plays an important role in information fusion applications. The assessment can validate the effectiveness of the fusion algorithm. Moreover, performance assessment results can be further incorporated in the algorithm to optimize or guide the fusion process. Although performance assessment is application dependent and the requirements may vary, there is a general rule that applies to many image fusion methods. Performance assessment for image fusion counts the amount of information transferred from the input images to the fused results. There are different ways to represent such transformation of information which have to be scored against desired results. For example, image fusion should not obscure targets seen in one image and not the other. Additionally, there are a number of choices in an overall fusion process, including variety of fusion algorithms, diversity of image data sets, and differences among performance assessment approaches, which make it a challenge to validate the final fused results. Generally, the fused results can be evaluated either subjectively or objectively.

1) Subjective assessment: Subjective evaluation incorporates a user's (or subject's) judgments on the quality of the fused image product. The user can vary their opinion based on mission needs, perceptual abilities, or personal feelings. To evaluate a subjective analysis for image fusion, a new data set is needed which includes subject studies. A first of a kind data set has been produced for image fusion subjective assessment [6]. Sixty-three subjects were organized to perform a semantic segmentation on both the input and fused images. A "gold reference" is generated from the inputs and serves as the ground truth for the comparison with the fused image. The ground truth is compiled from users determining the pixels of interest, e.g. where a target is, the image quality, and significant boundaries based on image contrast. The procedure to generate the gold reference is illustrated in Fig. 3.

According to [6], the subjects were first asked to divide each image into pieces, where each piece represents a distinguished thing in the image. The manual segmentations of images from individuals are combined into a reference that can be used to evaluate the fusion results. As shown in the flowchart in Fig. 3, the boundary map resulted from corresponding segmentation is converted into a boundary mask image. The exact square two-dimensional Euclidean distance transform of the boundary map is first calculated using a square 3×3 structuring element. Then, the derived distance image processed with a threshold operation to obtain the binary mask image. The mask images from all the subjects are summed and thresholded at a level corresponding to half the number of subjects contributing to the sum [6]. Thus, the obtained result is called a *consensus binary mask image*. The logical union of the consensus binary mask images from visual and infrared inputs gives the reference mask image. A skeleton image containing boundaries of interest is derived with morphological operations and serves as the reference contour image. This procedure is applied to the input and fused images respectively. The contour from fused image is then compared with the reference contour generated from input images. The comparison employed the precision-recall framework, where the precision (P) and recall (R) are defined as:

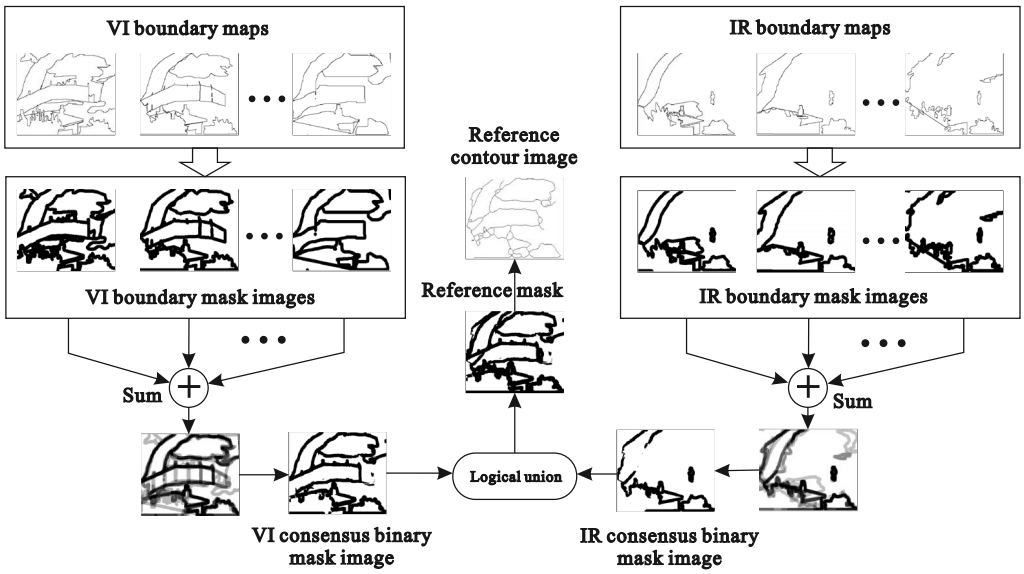


Fig. 3. Procedure to generate a "gold reference" from subjective assessment.

$$P = \frac{\text{number of correctly detected reference boundary pixels}}{\text{total number of detected boundary pixels}} \quad (1)$$

$$R = \frac{\text{number of corretly detected reference boundary pixels}}{\text{total number of reference boundary pixels}} \quad (2)$$

The F-measure is then defined as $F = 2PR/(R+P)$. A larger F-measure value between the fused result and "gold" reference indicates a better fusion performance.

2) Objective assessment: The computational model for image fusion objective assessment is also known as a fusion metric. A comparative study of the state of the art of image fusion metrics is presented in [2], of which twelve fusion metrics of four categories were compared for the same data set using different fusion algorithms. The four categories include information based metric, image feature based metric, structure similarity based metric, and human perception inspired fusion metric. A summation is given in Table I.

TABLE I. SUMMARY OF OBJECTIVE FUSION METRICS [2][1].

Objective Fusion Metrics		
Information theory based metric	Q_{MI}	Normalized mutual information
	Q_{TE}	Mutual information based on Tsallias entropy
	Q_{NICE}	Nonlinear correlation information entropy
Image feature based metric	Q_G	Gradient-based fusion metric
	Q_M	Image fusion metric based a multi-scale scheme
	Q_{SF}	Image fusion metric based on spatial frequency*
	Q_P	Image fusion metric based on phase congruency
Image structural similarity based metric	Q_s	Piella's metric
	Q_C	Cvejie's Metric
	Q_Y	Yang's Metric
Human perception inspired fusion metric	Q_{CV}	Chen-Varshney metric*
	Q_{CB}	Chen-Blum metric

[1]Metrics with * are not considered in the experiment.

The value of nine metrics is rounded to the range $[0, 1]$ and are considered in this study. These metrics have been applied to medical, satellite, and forest imagery [17], [18], [19]. A larger value means a better result.

III. ANALYSIS OF FUSION PERFORMANCE DATA WITH STATISTICAL TESTS

Statistical tests are adopted in this study to identify the performance differences between the fusion algorithms. These comparative test results provide statistical verification and validation of the fusion results. Actually, the comparison can be conducted in three different ways: (1) only two methods are compared; (2) one method is compared to all the others ($1 \times N$); and (3) all methods are compared to each other ($N \times N$) [8]. In the test, a null hypothesis (H_0) and an alternative hypothesis (H_1) are defined. The null hypothesis states that there is no effect or no difference, whereas the alternative hypothesis claims the presence of an effect or a difference between algorithms. As per the scientific method, you can not prove H_1, but can disprove H_0, i.e. there was no change in the image fusion results. A significance value α is used to determine at which level the hypothesis should be rejected [10].

Parametric tests assume the normality and equal variances of the data, which is not always satisfied in practice. Thus, non-parametric tests are needed to work with data that is not normal or of equal variance. In this study, the Wilcoxon signed ranks test and the Friedman test are used for a two method comparison and a multiple comparison ($N \times N$), respectively.

A. Wilcoxon signed ranks test

The Wilcoxon signed-ranks test is a non-parametric alternative to the paired t-test and ranks the differences in performances of two algorithms for each data set [8], [20], [10]. It compares the ranks for the positive and negative differences.

Let d_i be the difference between the F-measure values of two assessment approaches on i-th out of n problem or data set. The differences are ranked based on the absolute values. The use of average ranks is recommended to deal with ties. Let R^+ be the sum of ranks for the problems where the first assessment value is larger than the second. And R^- is the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored [10]. The sum of ranks are defined as:

$$R^+ = \sum_{d_i>0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i) \tag{3}$$

$$R^- = \sum_{d_i<0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i) \tag{4}$$

Let $T = \min(R^+, R^-)$ if T is the smaller sum. When T is less than or equal to the value of the distribution of Wilcoxon for n degrees of freedom, the null hypothesis of equality of mean is rejected. For a larger number of data sets, the statistic

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{5}$$

is distributed approximately normally. With $\alpha = 0.05$, the null hypothesis will be rejected if z is smaller than -1.96 [8]. In our case, the conclusion is that the two assessments are different (see section IV).

B. Friedman test with post hoc analysis

1) Friedman test: The Friedman test is a non-parametric equivalent of the repeated-measures Analysis of Variance (ANOVA) that determines the difference between group means [8]. The Friedman test carries out a multiple comparison test to detect significant differences between two or more algorithms [10]. The algorithms are ranked for each data set. In the case of ties, average ranks are assigned. The Friedman test compares the average ranks of algorithms with the null hypothesis that all the algorithms are equivalent or behave similarly [8].

To calculate the test statistic, the original results are first converted to ranks. The detailed procedures are as follows [10]:

- Collect results for each algorithm/problem pair;
- For each algorithm/problem i, rank values from 1 (best result) to k (worst result) as r_i^j $(1 \le j \le k)$; then
- Obtain the final rank $R_j = \frac{1}{n}\sum_i r_i^j$ for each algorithm j.

The best algorithm should have the rank of 1. The Friedman statistic χ_F^2 can be computed as:

$$\chi_F^2 = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{6}$$

which is distributed according to χ_F^2 with $k-1$ degrees of freedom, when n and k are big enough. For a smaller number of algorithms and data sets, exact critical values have been computed [10], [8]. Iman *et al.* proposed a better statistic, which avoids the undesirable conservative of χ_F^2 [21], [8], [10]. The proposed statistic is [10]:

$$F_{ID} = \frac{(n-1)\chi_F^2}{n(k-1) - \chi_F^2} \tag{7}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. The table of critical values can be found in statistical books.

2) Post hoc analysis: The Friedman test detects the significant difference over the complete multiple method (or group) comparisons, but it does not tell which group. Thus, it is necessary to determine which pairs in the group have the significant differences. The Nemenyi test is used when all the algorithms are compared to each other. The performance of the two algorithms is significantly different if the corresponding average ranks differ by at least the critical difference (CD), which is given as:

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6n}} \tag{8}$$

where $q_{\alpha=0.05,0.10}$ is based on the studentized range statistic divided by $\sqrt{2}$. The critical value is adjusted for making $k(k-1/2)$ comparisons. A table of critical values for the two-tailed Nemenyi test can be found in [8]. For a comparison of all algorithms, there will be $k(k-1)/2$ hypotheses and the post hoc analysis can obtain a p-value which determines the degree of rejection of each hypothesis.

The p-value of every hypothesis can be obtained through the conversion of the rankings by using a normal approximation. The test statistic for comparing the i-th and j-th algorithm is [10]:

$$z = (R_i - R_j)/\sqrt{\frac{k(k+1)}{6n}} \tag{9}$$

The z value is used to find the corresponding probability from the normal distribution table, which is then compared with an appropriate α [8]. The value of α needs to be adjusted to compensate for the multiple comparisons. The Nemenyi test calculates the adjusted value of α in a single step by dividing it by the number of comparisons, i.e. $k(k-1)/2$ [22]. Therefore, the adjusted p-value for Nemenyi is: $\min\{v; 1\}$, where $v = k(k-1)p_i/2$.

The statistical tests in the following experiments were conducted with R [23] and the KEEL tools [24].

IV. Experimental Results

The data sets used in the experiments contain image fusion assessment results by a group of subjects (user) and an expert (expert) [6]. The data included ten sets of multi-sensor images were fused by three MRA algorithms, i.e. PYR, DWT, and CWT. For the objective assessment, we considered the four categories of fusion metrics, including ten selected algorithms with metric values in the range $[0, 1]$ [2].

A. Test 1: User vs. expert assessments

The first experiment investigates the difference of the assessments between the user and the expert in terms of the F-measure. Boxplots of the F-measure for different fusion algorithms by the user and expert are given in Fig. 4(a) and 4(b). The boxplot for the expert's and user's assessments is also given in Fig. 4(c) for comparison. Both the median and mean (red point) values are highlighted. The data differences can be observed. The null hypothesis H_0 states that there is no difference between the assessments of the user and expert while the alternative hypothesis says that they are different. The Wilcoxon signed-ranks test was performed. Table II gives the results at the significance level $\alpha = 0.05$. Herein, VS refers to the sum of ranks assigned to the F-measure differences between the user and expert with positive (R^+) or negative (R^-) sign. The p-value 0.241 is larger than 0.05, which means the null hypothesis cannot be rejected, i.e. there is no significant difference between the assessments from the user and expert.

TABLE II. RESULTS OBTAINED BY THE WILCOXON SIGNED RANKS TEST AT A LEVEL OF SIGNIFICANCE $\alpha = 0.05$.

VS	R^+	R^-	p-value
User versus Expert	290.0	175.0	0.2410

The mean values of the F-measure are listed in Table III. The maximum values are highlighted. The expert prefers the PYR while user's favourite is CWT, when only mean value is considered.

TABLE III. THE MEAN VALUE OF F-MEASURE.

	PYR	DWT	CWT
User	0.53718	0.51063	**0.55032**
Expert	**0.60056**	0.56420	0.56329

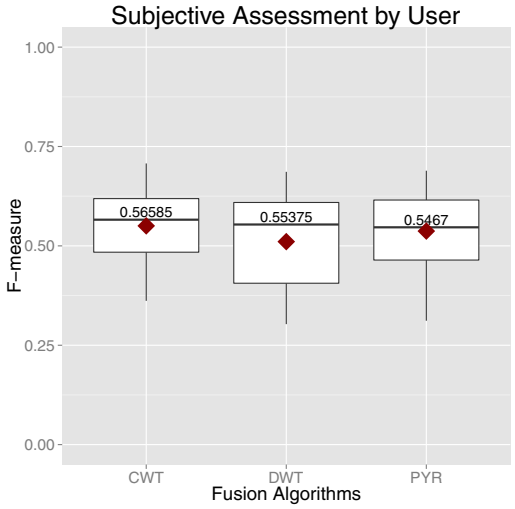
B. Test 2: PLIF method comparisons from user assessments

In the second experiment, multiple comparisons were carried out for different fusion algorithms based on the users' assessments. The Friedman test tells whether the three approaches are similar or not. The test results and ranks by user's assessment are given in Table IV. However, this rank does not give much information about the performance of the algorithms. Unlike the classification application, the classifiers are evaluated based on either the accuracy or the error rate. A higher rank refers to a higher accuracy or a lower error rate. Thus, it is an indication of the performance. For image fusion, the fusion performance value is evaluated. A higher rank can be reached by simply comparing the performance values. So, according to Table IV, a preference is given to PYR in terms of the average ranks from the Friedman test of users' assessments.

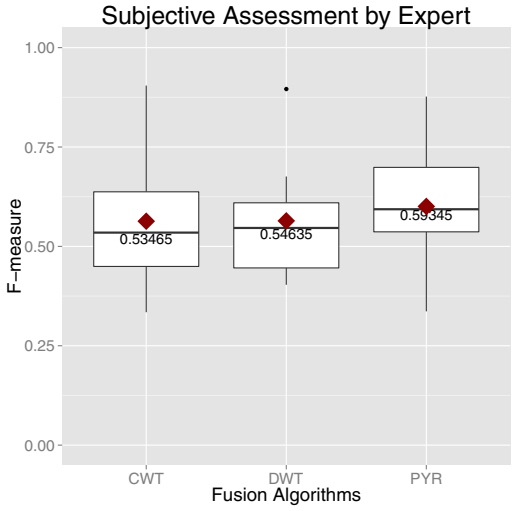
TABLE IV. THE FRIEDMAN TEST OF USER'S ASSESSMENTS.

Algorithm	User.PYR	User.DWT	User.CWT
Ranking	1.7	2.4	1.9
χ^2	2.6	degree of freedom	2
p-value		0.2725	

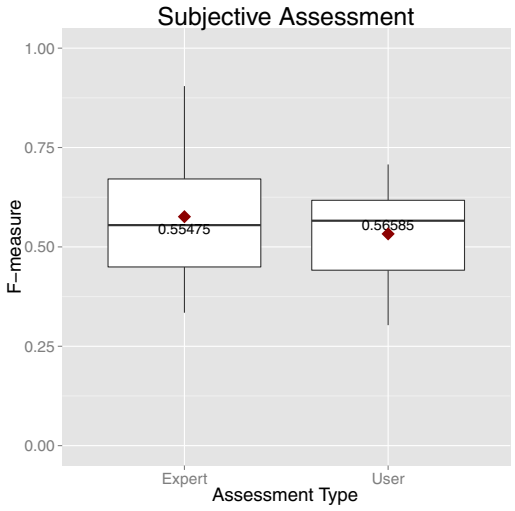
However, the p-value is larger than 0.05, the difference between these fusion algorithms are not significant in terms of users' assessments. As the difference is not statistically significant, there is no need to conduct the post hoc analysis,



(a)



(b)



(c)

Fig. 4. The subjective assessments by expert and user.

but the Nemenyi test results are listed in Table V for information. Nemenyi's procedure rejects those hypotheses that have an unadjusted p-value ≤ 0.016667. Thus, all the hypotheses cannot be rejected according to the p-value in Table V. The "major difference" is between the PYR and DWT by the users' assessments.

TABLE V. ADJUSTED p-VALUES FOR SUBJECTIVE COMPARISON OF FUSION ALGORITHMS (BY USER).

i	Hypothesis	Unadjusted p	p_{Neme}
1	User.PYR versus User.DWT	0.117525	0.352575
2	User.DWT versus User.CWT	0.263552	0.790657
3	User.PYR versus User.CWT	0.654721	1.964163

The test results on expert's assessments is given in Table VI. Again, there is no consequential difference observed at a significance level of 0.05. The difference between PYR and DWT is relatively more salient that the other pairs, although all are trivial in terms of expert's assessments (see Table VII).

TABLE VI. THE FRIEDMAN TEST OF EXPERT'S ASSESSMENTS.

Algorithm	Expert.PYR	Expert.DWT	Expert.CWT
Ranking	1.7	2.1	2.2
χ^2	1.4	degree of freedom	2
p-value		0.4966	

TABLE VII. ADJUSTED p-VALUES FOR SUBJECTIVE COMPARISON OF FUSION ALGORITHMS (BY EXPERT).

i	Hypothesis	Unadjusted p	p_{Neme}
1	Expert.PYR versus Expert.DWT	0.263552	0.790657
2	Expert.DWT versus Expert.CWT	0.371093	1.11328
3	Expert.PYR versus Expert.CWT	0.823063	2.46919

C. Test 3: Fusion metric comparisons

In the third part of the experiment, ten objective fusion metrics are considered. Figure 5 shows the metric values for three algorithms. The Friedman test ($N \times N$) is performed and a p-value$= 0.0247 < 0.05$ is obtained, which means there is a significant difference in the test. The CWT gains the highest rank 1.3 among the three. PYR and DWT are quite similar.

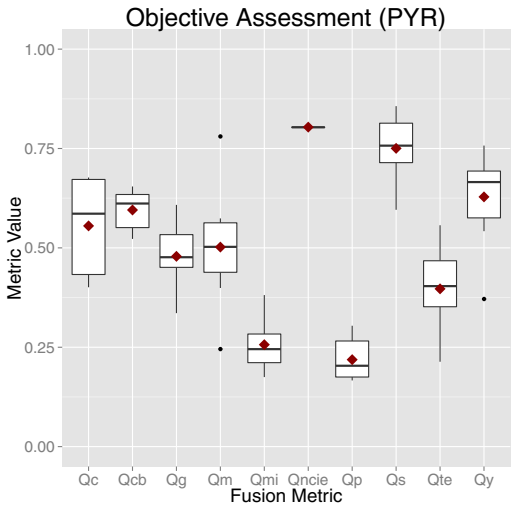
To understand where the differences happen, the Nemenyi test was conducted and results are given Table VIII. Nemenyi's procedure rejects those hypotheses that have an unadjusted p-value$= 0.013906 \leq 0.016667$ (see Table IX). Thus, the null hypothesis for DWT and CWT is rejected. In other words, the two algorithms are significantly different in the opinion of objective metrics while the differences among other pairs are trivial.

TABLE VIII. THE FRIEDMAN TEST OF OBJECTIVE FUSION METRICS.

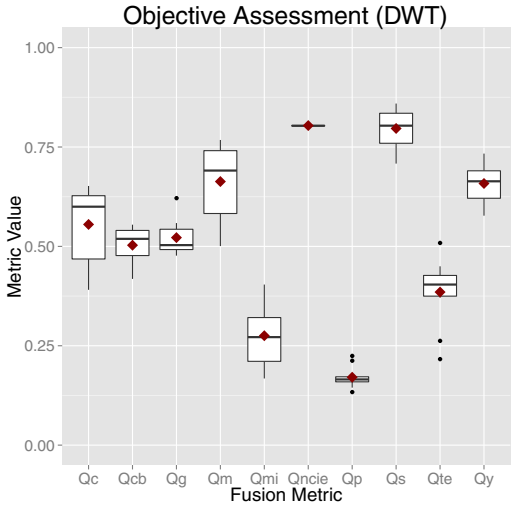
Algorithm	Objective.PYR	Objective.DWT	Objective.CWT
Ranking	2.3	2.4	1.3
χ^2	7.4	degree of freedom	2
p-value		0.0247	

TABLE IX. ADJUSTED p-VALUES FOR OBJECTIVE COMPARISON OF FUSION ALGORITHMS.

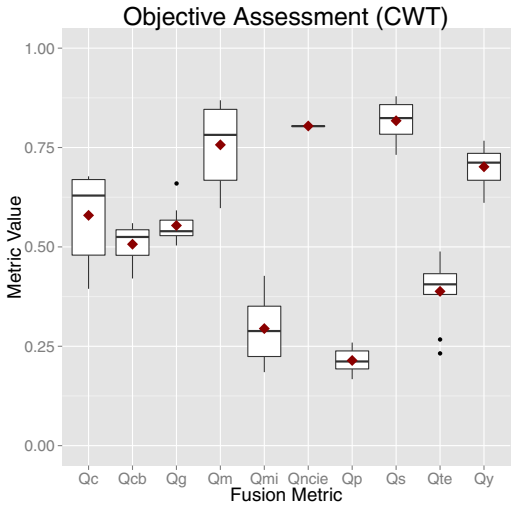
i	Hypothesis	Unadjusted p-value	p_{Neme}
1	DWT versus CWT	0.013906	0.041719
2	PYR versus CWT	0.025347	0.076042
3	PYR versus DWT	0.823063	2.46919



(a)



(b)



(c)

Fig. 5. The objective assessments with fusion metrics.

V. Discussion

The results differ between the user and the fusion methods. The assessments from the users and the expert are quite the same, i.e. no significant statistical difference is observed. This is understandable and reasonable as many of the images presented to these groups had already been enhanced for targeting. However, there is no strong evidence to support the significant difference between the three fusion algorithms from the user's assessment. In other words, the three fusion algorithms perform equally well or the human subjects could not tell the differences through the designed evaluation process. There are several potential explanations for user not being able to distinguish between methods. First, in the subjective assessment, the input images and fused result of the same set were assessed by different subjects. For example, the inputs of image set A are assessed by subject number one, but the fused images of set A are assessment by subject number two. This may introduce potential turbulence in the assessment results. Even though the "gold reference" is a combination of all subjects' assessments, it would be better to include a complete assessment a specific image set by the same subject. Second, the assessment is implemented through human semantic segmentation, which aims at identifying important pieces in the image and the results are user dependent. Thus, this process mimics a higher level perception of human being on object detection while some details in their selection process may be missed.

On the other hand, a set of objective metrics report the differences between the three fusion algorithms through the Friedman test and the major difference is highlighted by the Nemenyi test's result. The reason is that the objective metrics may capture the details in the image. The weights on such details highlight the differences between fusion algorithms. The phenomenon of differences between image fusion between the user and the computer still need further analysis. How to conduct the subjective assessment on pixel-level image fusion is still and open question and of concern of Level 5 Information Fusion [25]. Actually, the subject assessment depends on how the fused image is used, e.g. a visualization to the end user or further processing for a higher level understanding. Generally, the image fusion method should integrate as much as possible the details from input data such as collected metadata, pixel normalizations, and mission needs. Thus, both the objective and subjective assessment can benefit from contextual information to provide more meaningful results.

There are two scenarios of using statistical analysis to perform fusion assessment as illustrated in Fig. 6. The first is the assessment with subjective data. In an ideal case, all the algorithms for comparison including the proposed one (A_{new}) need to be assessed by subjects. Either pairwise or multiple comparisons ($1 \times N$ or $N \times N$) can be conducted with the assessment results. Whenever there is a new fusion algorithm proposed, the subjective assessment has to be repeated. This is costly and not practical. The second scenario is with the objective fusion metric. In this case, individual or multiple fusion metrics can be used. The conclusion could be the new fusion algorithm outperform algorithms A_1 and A_2 in terms of fusion metric B_1 or a combination of B_1 and B_2 etc. As the fusion metric may come with opposite results over different data sets, the statistical analysis can avoid such confusion during the comparison.

Some of the subjective results can be used to develop object score such as the National Image Interpretability Rating Scales (NIIRS) [26], [27]. When a new image fusion method is tested then a subjective-based objective score is determine. After a significance difference in the objective and pseudo-subjective metrics; then a user audit/pedigree trail can be conducted for further analysis.

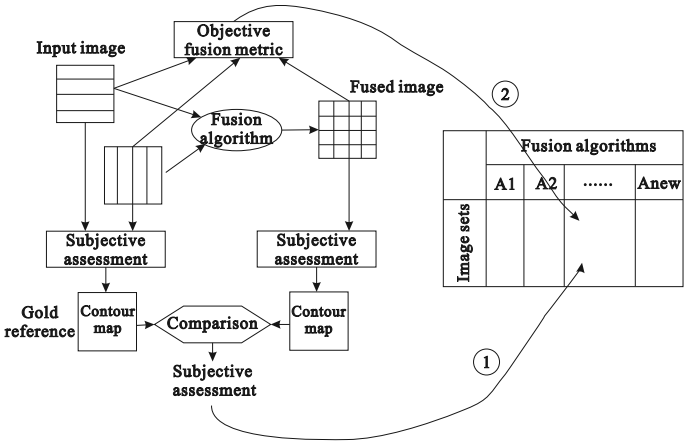


Fig. 6. The fusion performance assessment by statistical analysis.

To validate a new fusion algorithm, the comparison with other fusion algorithms is performed using a series of fusion metrics to do verification. However, the validation issue exits in the objective fusion metrics as well against the mission needs. A similar procedure with statistical analysis can be applied to compare the objective fusion metrics. When the subjective ground truth is available, we can learn which metric better matches the user needs. When there is no such reference, we can only tell the differences between those metrics; but not which is better. However, the fusion algorithm will also affect the fusion metric. Therefore, the fusion metric study needs to be carefully and clearly defined as well against the operating conditions, mission needs, user objectives, and control of the image fusion parameters.

VI. Concluding Remarks

This paper presents the use of statistical analysis for image fusion performance assessment. Non-parametric tests, i.e. Wilcoxon signed ranks test and Friedman test with post hoc analysis, are employed to evaluate three multi-resolution fusion algorithms. In the tests, the semantic segmentation based subjective assessment does not exhibit any significant differences between the three algorithms while the objective metrics highlight a significant difference. The contrasting results may be partly due to the subjective assessment method itself.

The statistical tests make it possible to compare fusion algorithms over multiple image data in terms of individual or multiple metrics. The statistical tests introduced here for image fusion can also be applied to validate the fusion metrics over multiple data sets and fusion algorithms.

Future work includes three areas. For the community and building on the International Society of Information Fusion

(ISIF) fused imagery [6], the benchmark data set with human subjective assessment will be refined and available for further analysis. Also, we seek methods for user-defined objectives scores and are developing a interpretability rating scale for image fusion results. Finally, we will induce quantifiable errors in the imagery, such as blurring, to determine the balance between the user experience and choice of image fusion methods to deal with challenging scenarios.

ACKNOWLEDGMENT

The authors would like to thank Dr. Alex Toet and his colleagues for preparing the multi-sensor data sets and making it available for research use.

REFERENCES

[1] R. S. Blum and Z. Liu, Eds., *Multi-sensor Image Fusion and Its Applications*, ser. Signal Processing and Communications. Taylor and Francis, 2005.

[2] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 94–109, 2012.

[3] B. Kahler and E. Blasch, "Sensor management fusion using operating conditions," in *Proceedings of National Aerospace and Electronics Conference*, Fairborn, Ohio, USA, July 2008.

[4] E. Blasch, K. Laskey, A.-L. Jousselme, V. Dragos, P. Costa, and J. Dezert, "Urref reliability versus credibility in information fusion (stanag 2511)," in *Information Fusion (FUSION), 2013 16th International Conference on*, July 2013, pp. 1600–1607.

[5] E. Blasch, E. Bosse, and D. Lambert, Eds., *High-Level Information Fusion Management and Systems Design*, 1st ed. Artech House, Inc., April 2012, iSBN-13: 978-1-60807-151-7.

[6] A. Toet, M. Hogervorst, S. Nikolov, J. Lewis, T. Dixon, D. Bull, and C. Canagarajah, "Towards cognitive image fusion," *Information Fusion*, vol. 11, no. 2, pp. 95 – 113, 2010.

[7] Y. Zheng, W. Dong, and E. P. Blasch, "Qualitative and quantitative comparisons of multispectral night vision colorization techniques," *Optical Engineering*, vol. 51, no. 8, pp. 087 004–1–087 004–16, 2012.

[8] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, Dec. 2006.

[9] S. Garcia, A. Fernandez, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Information Sciences*, vol. 180, no. 10, pp. 2044 – 2064, 2010, special Issue on Intelligent Distributed Information Systems.

[10] J. Derrac, S. Garca, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3 – 18, 2011.

[11] S. Garcia, A. Fernndez, J. Luengo, and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability," *Soft Computing*, vol. 13, no. 10, pp. 959–977, 2009.

[12] J. Luengo, S. Garcia, and F. Herrera, "A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7798 – 7808, 2009.

[13] G. Piella, "A general framework for multiresolution image fusion: from pixels to regions," *Information Fusion*, vol. 4, no. 4, pp. 259–280, December 2003.

[14] 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*, vol. 87, no. 8, pp. 1315–1326, Aug 1999.

[15] J. J. Lewis, R. J. OCallaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Information Fusion*, vol. 8, no. 2, pp. 119 – 130, 2007, special Issue on Image Fusion: Advances in the State of the Art.

[16] "Image fusion website," http://www.imagefusion.org, August 2009.

[17] R. Shen, I. Cheng, and A. Basu, "Cross-scale coefficient selection for volumetric medical image fusion," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 4, pp. 1069–1079, April 2013.

[18] E. P. Blasch and Z. Liu, "LANDSAT satellite image fusion metric assessment," in *Proceedings of the 2011 IEEE National Aerospace and Electronics Conference (NAECON)*, July 2011, pp. 237 –244.

[19] X. Ding, L. Yan, J. Liu, J. Kong, and Z. Yu, "Obstacles detection algorithm in forest based on multi-sensor data fusion." *Journal of Multimedia*, vol. 8, no. 6, pp. 790–795, 2013.

[20] P. Dalgaard, *Introductory Statistics with R*, 2nd ed., ser. Statistics and Computing, J. Chambers, D. Hand, and W. Hardle, Eds. New York, USA: Springer, 2008.

[21] R. L. Iman and J. M. Davenport, "Approximations of the critical region of the Friedman statistic," *Communications in Statistics - Theory and Methods*, vol. 9, no. 6, pp. 571–595, 1980.

[22] B. Trawinski, M. Smetek, Z. Telec, and T. Lasota, "Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms," *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 4, pp. 867–881, December 2012.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org

[24] http://www.keel.es/, retrieved 2014.

[25] E. Blasch, *Handbook of Multisensor Data Fusion: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2008, ch. Introduction to Level 5 Fusion: the Role of the User, pp. 503–535.

[26] Intelligence Resouce Program, "National image interpretability rating scales," https://www.fas.org/irp/imint/niirs.htm, retrieved 2014.

[27] B. Kahler and E. Blasch, "Predicted radar/optical feature fusion gains for target identification," in *Aerospace and Electronics Conference (NAECON), Proceedings of the IEEE 2010 National*, July 2010, pp. 405–412.