Information Fusion 36 (2017) 251–260

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Statistical comparison of image fusion algorithms: Recommendations

Z. Liu[a,*], E. Blasch[b], V. John[c]

[a] School of Engineering, Faculty of Applied Science, University of British Columbia (Okanagan), 1137 Alumni Avenue Kelowna, BC V1V 1V7 Canada
[b] Information Directorate, Air Force Research Laboratory, Rome, NY, 13441 USA
[c] Toyota Technological Institute, Tenpaku-Ku, Nagoya, 468-8511 Japan

ARTICLE INFO

Article history:
Received 19 May 2016
Revised 19 December 2016
Accepted 20 December 2016
Available online 23 December 2016

Keywords:
Image fusion
Pixel-level fusion
Performance assessment
Statistical comparison
Statistical significance

ABSTRACT

Pixel-level image fusion has been applied in a variety of applications, including multi-modal medical imaging, remote sensing, industrial inspection, video surveillance, and night vision etc. Various algorithms are being proposed for numerous applications which requires a comprehensive method of assessment to discern which methods provide decision support. Currently, the validation or assessment of newly proposed algorithms is done either subjectively or objectively. A subjective assessment is costly and affected by a number of factors that are difficult to control. On the other hand, an objective assessment is carried out with a fusion performance metric which is defined to evaluate the effectiveness and/or efficiency of the fusion operation. There are a number of fusion metrics proposed for fusion processes taking different perspectives. Most image fusion research presents a comparison of the proposed and existing fusion algorithms with selected fusion metric(s) over multiple image data sets. The proposed algorithm advantage is justified by the relative difference with the best or better metric values. However, the statistical significance of such difference is unknown leading to a misperception of the quantitative differences between methods. This paper proposes the use of non-parametric statistical analysis for comparisons of fusion algorithms along with the Image fusion Toolbox Employing Significance Testing (ImTEST). Strategies to use different tests in varied scenarios are presented and recommended. Experiments with recently published algorithms demonstrate the necessity to adopt the statistical comparison to establish a baseline for image fusion research.

1. Introduction

Various algorithms have been proposed to implement pixel-level image fusion, which has found a diverse range of applications [1]. A performance assessment needs to be conducted with fusion performance metrics to validate the effectiveness as well as the significance of the proposed algorithms. A comprehensive assessment will indicate which fusion algorithm performs better for a specific set of sensors, targets, and environment data and applications [2,3]. However, the accuracy and reliability of the fusion metric is not complete [4]. For example, each metric may reveal one inherent property of the fusion process or the fused image, but not a comprehensive performance over all operating conditions. Multiple metrics may give comprehensive but contrary judgments. Thus, it is necessary to establish a baseline evaluation for the image fusion performance based on both absolute and relative comparisons.

In a subjective assessment study carried out by European researchers, a precision-recall framework was proposed based on the semantic segmentation of fused and input images [5]. Toet et al. described a procedure to establish a ground truth reference with the segments to assess fused results [5]. However, subjective assessment is costly and how to use the subjective assessment results has not been fully exploited as a framework for image fusion comparisons (i.e., subjective assessments are subject to individual differences). Most image fusion research employs a number of fusion performance metrics to conduct comparisons with selected fusion algorithms [6]. When most metrics indicate a “better” value, e.g. from 0.9534 to 0.9589, the research claims a new contribution or improvement. However, the significance of the “improvement” cannot be confirmed by simply comparing the value.

Statistical comparisons have been applied to the data mining research to evaluate a newly proposed algorithm [7–11]. Different classifiers can be compared with a set of non-parametric statistical tests. With the tests, it is possible to determine which classification algorithm is considered better than another as well as the performance gain. Likewise, when a new algorithm is proposed, it

[*] Corresponding author.
E-mail addresses: zheng.liu@ieee.org (Z. Liu), erik.blasch.1@us.af.mil (E. Blasch), vijayjohn@toyota-ti.ac.jp (V. John).

http://dx.doi.org/10.1016/j.inffus.2016.12.007
1566-2535/© 2016 Elsevier B.V. All rights reserved.

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 a variety of specific conditions. In our pilot study, we used the data sets presented in [5] with the non-parametric statistical tests on subjective assessment and objective fusion metrics. The test results lead to some discussions about the understanding of subjective assessment and the use of fusion metrics. The Nemenyi test was employed in our previous work for the *post-hoc* analysis, in which a fixed critical difference value was used. However, the Nemenyi test is conservative with less statistical power. Thus, this study employed different *post-hoc* procedures based on the test methods and the number of algorithms in the multiple comparisons.

This paper proposes the use of statistical comparison method to assess the performance of image fusion algorithms. The proposed method will help to establish a comparative procedure to evaluate fusion algorithms. Three case studies with recently published data are carried out in this study to confirm the feasibility of the proposed approaches. The rest of the paper is organized as follows. Section 2 briefly describes the pixel-level image fusion and fusion performance assessment. The use of statistical tests for algorithm comparison is explained in Section 3. Experimental results with three recently proposed fusion algorithms are presented and discussed in Section 4. Section 5 concludes this paper.

2. Pixel-level image fusion and performance assessment

Pixel-level image fusion generates a composite image from multiple sources to obtain richer content than any inputs. There are two major image fusion frameworks, i.e. multi-resolution analysis (MRA) and sparse representation (SR). The basic idea is to represent the source images with some predefined or learned basis functions because salient image features can be highlighted and easy to manipulate through such representations or coefficients. The combination operation is then applied to integrate those representations or coefficients. The corresponding inverse operation obtains the fused image from the combined representations or coefficients. The flexibility of MRA-based fusion comes from inherent properties of the MRA basis functions, such as overcompleteness, orthogonality, biorthogonality, rotation-invariance, and translation-invariance etc. The SR-based fusion benefits from the basis functions, which are known as atoms, represented in a dictionary, which is created or learned from the training data sets. Both frameworks need some combination rules to integrate those coefficients or representations. In [12], the MRA and SR methods are combined in a general framework for image fusion. Besides the MRA and learned dictionaries, other transforms, which represent image in the corresponding transform domain, have been employed as well, such as fractional Fourier transform [13], guided filtering [14], higher order singular value decomposition [15], compressive sensing [16], and empirical mode decomposition [17] etc. The use of artificial neural networks was reported in [18], where a probabilistic neural network and a radial basis function network were used to select the input image blocks based on designated features. A cooperative neural fusion regularization algorithm was proposed in [19], which can achieve the optimal image estimate with the less loss of contrast information. Other approaches include fusion in color space [20–23], adaptive fusion [24], optimization-based approaches [25–29], and cascade fusion [30]. Assembling varied elements in the fusion frameworks offers different options to implement image fusion. However, significant improvements have not been elucidated from these types of work. Thus, the image fusion research needs an assessment to validate such significance in fusion performance.

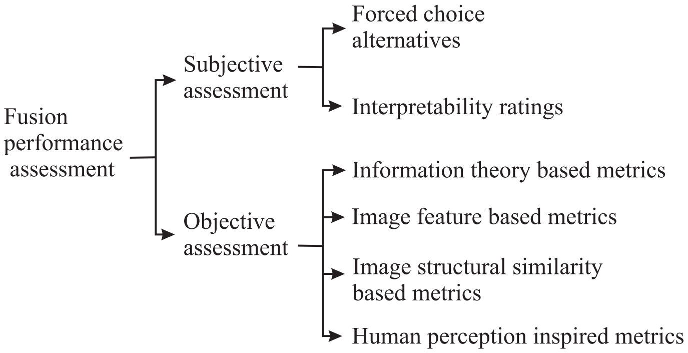

Fusion performance assessment
Subjective assessment
Forced choice alternatives
Interpretability ratings
Objective assessment
Information theory based metrics
Image feature based metrics
Image structural similarity based metrics
Human perception inspired metrics

Fig. 1. Methods for image fusion performance assessment.

Generally, the methods to assess image fusion performance consist of subjective and objective categories as shown in Fig. 1. Subjective assessment needs to involve human subjects, which is costly and difficult to handle the variances associated with biological differences, cognitive abilities, user preferences. Objective assessment relies on a computer model using a “fusion metric” to generate a numeric value to rate the fusion operation. There are basically four classes of objective metrics: information theory based metric, image feature based metric, image structural similarity based metric, and human perception inspired metric. [6]. A comprehensive description of these metrics is available in [6]. Each metric represents one aspect of the fusion operation process. All the current metrics focus on the “process” rather than the “result”. In other words, these metrics measure the goodness of the fusion operation in terms of signal-to-noise ratio, entropy, and information or image content, etc. It is still not clear how the fused image benefits the intended context of a specific application. The use of the fusion metrics is sometimes problematic as slight changes in the metric values do not always justify a significant improvement by the fusion algorithm. Thus, a statistical analysis as presented and recommended in the rest of this paper.

3. Statistical comparison

Statistical tests are adopted for multiple algorithm comparisons, which can provide statistical verification and validation of the fusion results. In a 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. Using the scientific method, the strategy is to determine if the null hypothesis is false (i.e., there is a significant difference between the methods). A significance value α is used to determine at which level the hypothesis should be rejected [9]. A larger α will ensure not missing detecting a difference that may exit while a smaller α is to ensure detecting a difference that really does exit. However, the smaller the significance level, the less likely to make a Type I error (rejecting a true null hypothesis), and the more likely to make a Type II error (failing to reject a false null hypothesis). Thus, an appropriate significance level should be chosen to balance these opposing risks of erros. Usually, $\alpha = 0.05$ is selected, which indicates a 5% risk of concluding that a difference exists when there is no actual difference. 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 unequal variance. In this study, the Wilcoxon signed ranks test and the Friedman test are used for

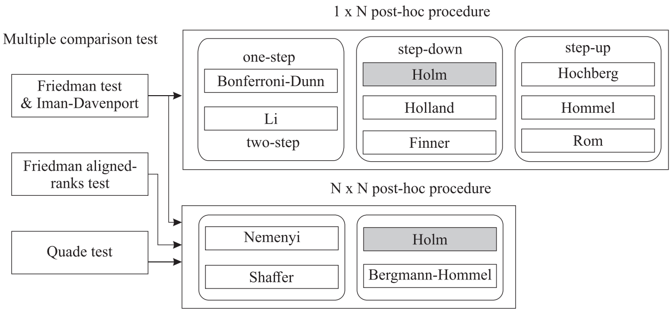


Fig. 2. Summary of nonparametric tests and post-hoc procedures.

a two method comparison (1 × 1) and a multiple comparison ($N \times N$), respectively.

3.1. Comparison of two algorithms

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 [7,9,31]. It compares the ranks for the positive and negative differences.

Let d_i be the difference between the fusion metric values of two fusion methods on i-th out of n problems or data sets. 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 fusion metric 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 [9]. The sum of ranks are defined as:

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

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

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

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{3}$$

is distributed approximately normally. With $\alpha = 0.05$, the null hypothesis will be rejected if z is smaller than -1.96 [7]. In this application, the conclusion is that the two fusion algorithms are different.

3.2. Comparison of multiple algorithms

Multiple algorithm comparison can be conducted with a control method ($1 \times N$) and among all methods ($N \times N$). Fig. 2 illustrates the multiple comparison nonparametric tests as well as corresponding *post-hoc* procedures. First, the Friedman test and/or its two alternatives, Friedman aligned ranks test and Quade test, are performed to detect whether statistically significant differences exist among the fusion algorithms [32]. If statistical significance is present, a corresponding *post-hoc* procedure should be carried out to identify which pair of algorithms differs significantly.

3.2.1. 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 [7]. In contrast to the common ANOVA, there is no need to assume that all the samples are from normal distribution in the Friedman test. It carries out a multiple comparison test to detect significant differences between two or more algorithms [9]. 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 [7].

To calculate the test statistic, the original results are first converted to ranks. Let r_i^j be the rank of the j-th of k algorithms on the i-th of n data sets. The detailed procedures are as follows [9]:

- 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 null-hypothesis states that all the algorithms are equivalent and the corresponding ranks R_j should be equal. The Friedman statistic χ_F^2 can be computed as:

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

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 [7,9]. Iman et al. proposed a better statistic, which avoids the undesirable conservative of χ_F^2 [33]. The proposed statistic is [9]:

$$F_{ID} = \frac{(n-1)\chi_F^2}{n(k-1) - \chi_F^2} \tag{5}$$

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 like [34].

Detailed descriptions of Friedman aligned ranks test and Quade test are available in [9]. They output quite similar results but they are more powerful for no more than five algorithms. The Friedman test with Iman and Davenport extension is a good choice for more than five different algorithms [35].

3.2.2. 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. A set of hypotheses can be defined for ($1 \times N$) and ($N \times N$) respectively. In ($1 \times N$) comparisons, the selected control algorithm is compared with the rest $k-1$ algorithms. The ($N \times N$) comparisons consider $k(k-1)/2$ comparisons among algorithms. The *post-hoc* procedure derives 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. For the Friedman test, the statistic z for comparing the i-th and j-th algorithm is [9]:

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

The z value is used to find the corresponding probability from the normal distribution table, which is then compared with an appropriate α [7]. The value of α needs to be adjusted to compensate for the multiple comparisons. The definitions of z for the Friedman aligned ranks test and Quade test are given below. For Friedman aligned test, z is [9]:

$$z = (\hat{R}_i - \hat{R}_j)/\sqrt{\frac{k(n+1)}{6}} \tag{7}$$

where $\hat{R}_i$ and $\hat{R}_j$ are the average rankings by the Friedman aligned ranks test of the algorithms compared. For Quade test, z is defined as [9]:

$$z = (T_i - T_j)/\sqrt{\frac{k(k+1)(2n+1)(k-1)}{18n(n+1)}} \tag{8}$$

where there are $T_i = \frac{W_i}{n(n+1)/2}$ and $T_j = \frac{W_i}{n(n+1)/2}$. W_i and W_j are the rankings without average adjusting by the Quade test of the algorithms compared.

The obtained p-values need to be corrected to take into account the probability error of a certain comparison as well as the remaining comparisons. Fig. 2 also lists four classes of correction methods for the p-values, which include one-step, two-step, step-up and step-down methods. Detailed equations as well as for ($N \times N$) comparisons can be found in [9] and will not be duplicated here.

3.3. Procedure for fusion algorithm comparison

The suggested procedure for fusion algorithm comparison is illustrated in Fig. 3. Starting with counting the number of algorithms, the procedure ends up with a conclusion of “with improvement” or “without improvement” for the proposed new fusion algorithm. When there are only two algorithms, Wilcoxon signed-ranks test can be applied. When more than two algorithms are considered in the comparison, a statistical test needs to be carried out to detect if at least one algorithm performs different from the rest. Friedman test with Iman and Davenport extensions is often used to test more than five algorithms while Friedman aligned ranks and Quade test are employed if there are less than five to compare.

If such a significant difference is confirmed, a pair-wise test with corresponding *post-hoc* correction is applied for multiple comparisons. As described in Section 3.2.1, there are two types of comparison: comparison with a control or reference ($1 \times N$) and all pair-wise comparison ($N \times N$). To compare multiple algorithms with a control, the Friedman test with Finner’s correction is recommended as Finner is a simple but powerful procedure [35]. When an all pair-wise comparison is conducted, the choice of the *post-hoc* procedure to use with the Friedman test is based on the number of algorithms. The test for more than nine algorithms needs Shaffer’s static method or Finner and Holm procedures. Otherwise, the Bergmann and Hommel procedure are preferred [35].

When there are multiple data sets for each combination of image groups and fusion metrics, the mean value for each combination should be calculated and used in the Wilcoxon signed rank test with Finner’s correction. This case is needed to count image groups, which consist of multiple image sets. However, simply using images in groups may reduce the experimental data and is not recommended for statistical comparison.

4. Experimental results

Three case studies were carried out in the experiments. These cases are selected from recent publications [14,36,37]. For each case, the authors compared their proposed method with one or multiple fusion algorithms with selected fusion metrics. The judgment is based on the metric values. When the proposed method achieves a better result for most metrics, the effectiveness of the new approach is claimed.

Table 1
Comparison of results obtained from two compressive sensing based fusion algorithms (modified from Table 1 in [36]).

Image	Index	CS.MAV	CS.SD
1	MI	1.13	2.02
1	ENL	6.80	12.41
1	CC	0.49	0.58
2	MI	1.05	1.23
2	ENL	61.15	175.21
2	CC	0.68	0.75
3	MI	1.14	1.56
3	ENL	49.96	86.18
3	CC	0.72	0.76

For the experiments, we used a R package called “scmamp”, which is designed for statistical comparison of multiple algorithms in multiple problems [35]. A alternative tool with GUI is KEEL [38], which we used in our previous research [39]. Additionally, STAC provides a web service to perform statistical analysis of multiple algorithms [40]. The following analysis is reproducible by using our R code and data sets, which are available upon request and could serve as a sample implementation for the fusion performance assessment. The abbreviations of the fusion algorithms and fusion metrics are listed in the Appendices. Relevant references are also provided.

4.1. Results

4.1.1. Case study one

The first case is simple and selected from [36], where a fusion method based on compressive sensing is proposed. The original data set is given in Table 1, where three metrics are used for comparison with another compressive sensing based fusion algorithm [36]. As only two algorithms are involved, Wilcoxon signed rank test was conducted and resulted in a p-value 0.003843, which is smaller than 0.05. Thus, the difference of the two fusion algorithms is significant. Meanwhile, each fusion metric is considered respectively and Wilcoxon signed rank test was applied to each metric. The test gave the number 0.0544 for all the three metrics. As $0.0544 > 0.05$, the experiment is not able to claim a significant difference of the two method in terms of a specific metric. This raises a confliction with the overall test result. The reason is that there is not sufficient data for each metric as only three pairs of images were tested with each metric. The result for each group is presented as a single metric value. Therefore, sufficient data should be collected for a statistical comparison purpose.

4.1.2. Case study two

The second case is from [14], where guided filtering based fusion (GFF) is proposed as a new fusion algorithm. The original data are given in Table 2. There are total 8 algorithms compared with 3 groups of images and 5 fusion metrics (see Appendices for the details). Referring to the procedure in Fig. 3, the Friedman test with the Iman and Davenport extension (for more than five algorithms) was conducted and a p-value $2.2e-16$ was obtained. The p-value implies significant differences among the fusion algorithms. Further tests need to be conducted to investigate where the differences come from.

The first is a ($1 \times N$) Friedman test with Finner’s correction. The results are presented in Table 3. Each algorithm is compared with the proposed GFF algorithm serving as the control in comparison. The average ranking and the corrected p-value are listed in the table. The star (*) on the 1st row indicates the significant difference and this can be observed from the corrected p-value on the 2nd row, which is less than 0.05. The bold font highlights the

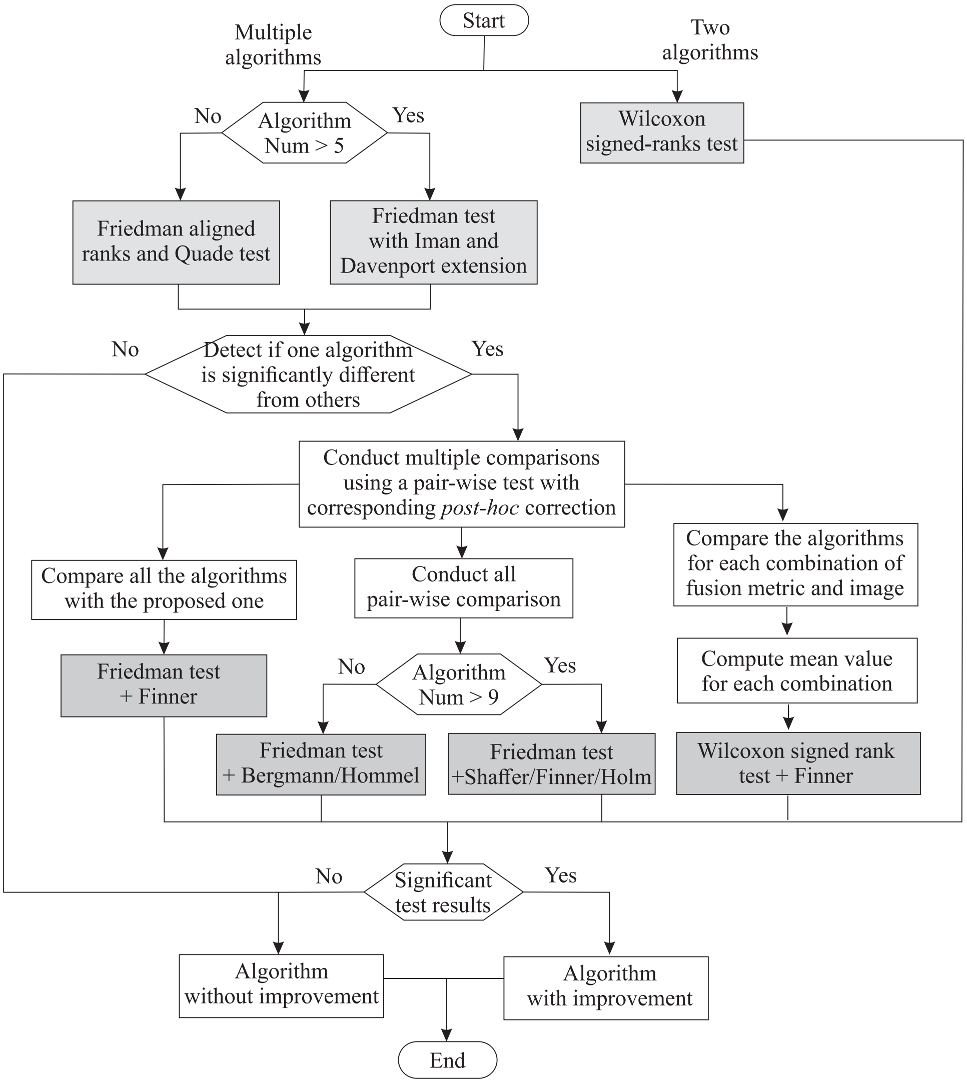

Start
Multiple algorithms
Two algorithms
Algorithm Num > 5
No
Yes
Wilcoxon signed-ranks test
Friedman aligned ranks and Quade test
Friedman test with Iman and Davenport extension
Detect if one algorithm is significantly different from others
No
Yes
Conduct multiple comparisons using a pair-wise test with corresponding post-hoc correction
Compare all the algorithms with the proposed one
Conduct all pair-wise comparison
Compare the algorithms for each combination of fusion metric and image
Friedman test + Finner
Algorithm Num > 9
No
Yes
Compute mean value for each combination
Friedman test + Bergmann/Hommel
Friedman test +Shaffer/Finner/Holm
Wilcoxon signed rank test + Finner
Significant test results
No
Yes
Algorithm without improvement
Algorithm with improvement
End

Fig. 3. The use of statistical tests for image fusion algorithm comparison.

Table 2
Quantitative performance assessment for different fusion methods (modified from Table I in [14]). GFF is the proposed fusion algorithm.

Image database	Metric	SWT	CVT	LAP	NSCT	GRW	WSSM	HOSVD	GFF
1	Q_y	0.862	0.813	0.868	0.864	0.696	0.809	0.967	0.934
1	Q_c	0.745	0.724	0.744	0.751	0.645	0.708	0.691	0.804
1	Q_g	0.632	0.56	0.644	0.633	0.446	0.617	0.648	0.657
1	Q_p	0.525	0.439	0.516	0.51	0.355	0.347	0.628	0.594
1	Q_{mi}	0.391	0.38	0.398	0.39	0.383	0.71	0.91	0.57
2	Q_y	0.915	0.894	0.922	0.911	0.761	0.877	0.955	0.964
2	Q_c	0.818	0.798	0.816	0.829	0.724	0.779	0.847	0.835
2	Q_g	0.681	0.661	0.698	0.673	0.519	0.668	0.685	0.714
2	Q_p	0.734	0.721	0.772	0.744	0.559	0.698	0.74	0.801
2	Q_{mi}	0.849	0.814	0.904	0.84	0.778	0.865	1.063	0.953
3	Q_y	0.717	0.738	0.792	0.798	0.717	0.827	0.953	0.914
3	Q_c	0.648	0.674	0.695	0.715	0.674	0.741	0.764	0.801
3	Q_g	0.605	0.575	0.693	0.672	0.474	0.638	0.62	0.704
3	Q_p	0.54	0.501	0.602	0.588	0.439	0.362	0.551	0.661
3	Q_{mi}	0.509	0.538	0.542	0.542	0.552	0.755	1.015	0.597

Table 3
($1 \times N$) test (Friedman test + Finner's correction) for the data in Table 2.

Fusion algorithm	SWT	CVT	LAP	NSCT	GRW	WSSM	HOSVD	GFF
Averaged ranking	5.1667*	6.4333*	3.5000*	4.1667*	7.4000*	5.3333*	2.4000	**1.6000**
Corrected p-value	0.0001167741	2.283135e−07	0.03914491	0.005748772	6.227849e−10	6.983672e−05	0.3710934	n/a

Table 4
Corrected p-values from ($N \times N$) pair-wise comparison with algorithm ranks in case two.

	SWT	CVT	LAP	NSCT	GRW	WSSM	HOSVD	GFF
SWT	n/a	0.784	0.323	0.961	0.113	1.000	**0.018**	**0.001**
CVT	0.784	n/a	**0.011**	0.101	1.000	0.961	**0.000**	**0.000**
LAP	0.323	**0.011**	n/a	1.000	**0.000**	0.323	0.961	0.236
NSCT	0.961	0.101	1.000	n/a	**0.004**	0.961	0.323	**0.049**
GRW	0.113	1.000	**0.000**	**0.004**	n/a	0.146	**0.000**	**0.000**
WSSM	1.000	0.961	0.323	0.961	0.146	n/a	**0.011**	**0.000**
HOSVD	**0.018**	**0.000**	0.961	0.323	**0.000**	**0.011**	n/a	1.000
GFF	**0.001**	**0.000**	0.236	**0.049**	**0.000**	**0.000**	1.000	n/a

Table 5
($1 \times N$) test (SWT removed) for the data in Table 2 (p-value$< 2.2e - 16$).

Fusion algorithm	CVT	LAP	NSCT	GRW	WSSM	HOSVD	GFF
Averaged ranking	5.6333*	3.3000*	3.8333*	6.5667*	4.7333*	2.3333	**1.6000**
Corrected p-value	9.503239e−07	0.03726332	0.006946538	1.827835e−09	0.0001424038	0.3525421	NA

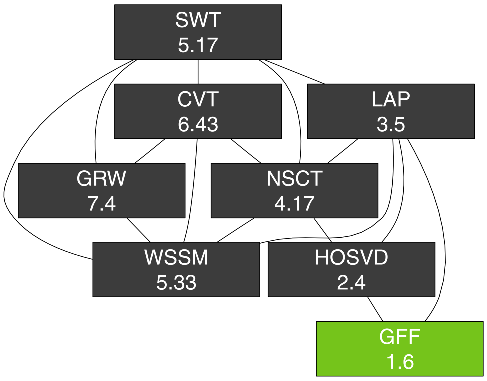


Fig. 4. Algorithm graph for all pair-wise comparison in case two.

best ranking (smaller is better). From this table, there is no significant difference observed between HOSVD and GFF methods, although these two are different from the rest.

The next is the ($N \times N$) pair-wise comparison done with the Friedman test with the Bergmann procedure (correction) as suggested in [35] for less than nine algorithms. The results are given in Table 4. Significant differences are highlighted with bold fonts. The proposed algorithm GFF is quite similar to algorithm LAP and HOSVD. More specifically, no obvious difference is observed between HOSVD and GFF. The result can also be represented with an intuitive graph based on corrected p-value, where connected nodes show no significant differences as illustrated in Fig. 4. Algorithm GFF, which has the highest ranking, is connected with LAP and HOSVD. The connection means that there is no significant difference between them. Fig. 5 presents the correlation matrix plot between the algorithms, which is directly calculated from fusion metric values. The cross over the number indicates that the correlation is insignificant.

In the first cast study, the original experiments employed only three pairs of images (nine samples in total). In the second one, three image databases were employed, but the results were presented with the averaged metric value for each database. The number of images, i.e. the sample size, does have impact on the power of statistical analysis, given the number of fusion algorithms. Further tests were conducted to see potential effect from sample size. Referring to Table 3, we sequentially removed the five algorithms, e.g. SWT, CVT, LAP, NSCT, and GRW, and repeated the statistical tests. The results of ($1 \times N$) test and ($N \times N$) comparison are given in Tables 5–14, respectively. The corresponding algorithm graphs are shown in Fig. 6.

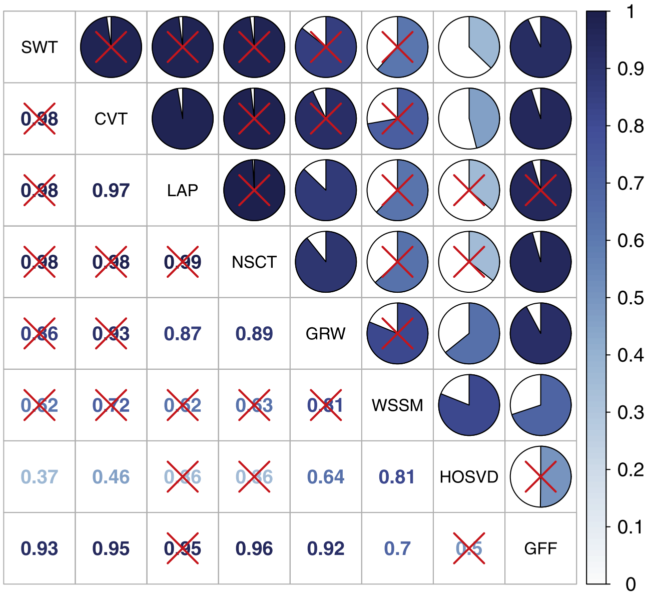


Fig. 5. The plot of the correlation matrix of the algorithms in case two. The cross indicates the correlation is not significant.

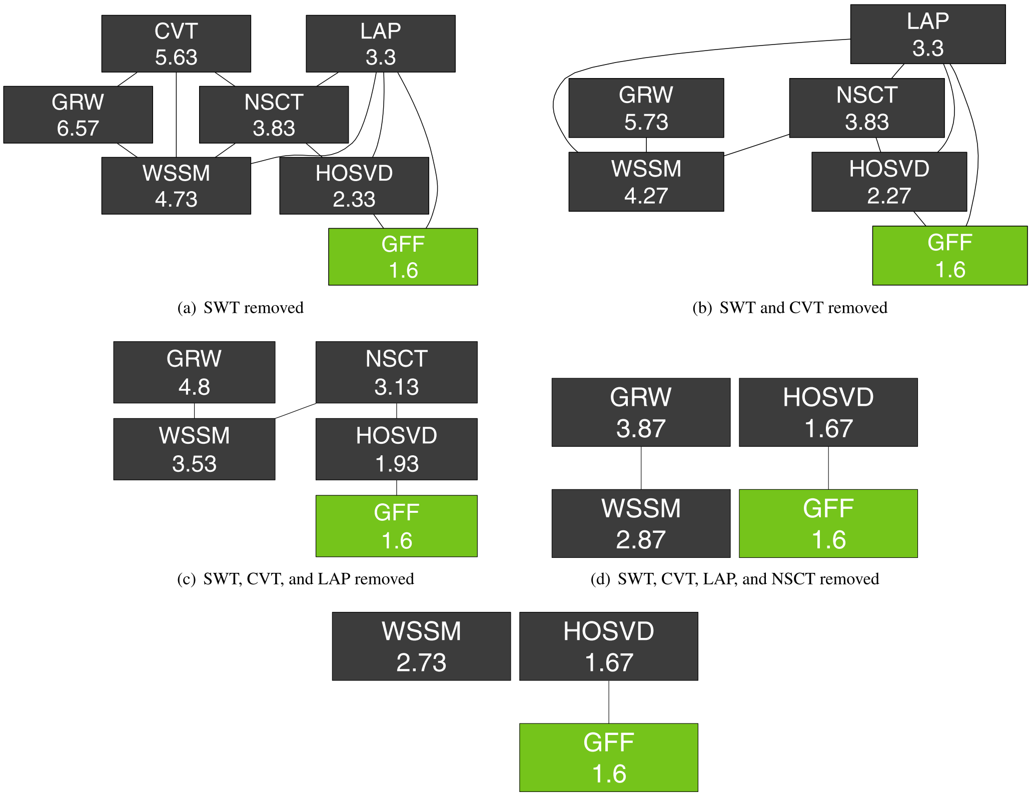
Table 6
Corrected p-values from ($N \times N$) pair-wise comparison (SWT removed) with algorithm ranks in case two.

	CVT	LAP	NSCT	GRW	WSSM	HOSVD	GFF
CVT	n/a	**0.022**	0.135	0.710	0.762	**0.000**	**0.000**
LAP	**0.022**	n/a	0.762	**0.000**	0.346	0.661	0.156
NSCT	0.135	0.762	n/a	**0.005**	0.762	0.229	**0.042**
GRW	0.710	**0.000**	**0.005**	n/a	0.121	**0.000**	**0.000**
WSSM	0.762	0.346	0.762	0.121	n/a	**0.016**	**0.001**
HOSVD	**0.000**	0.661	0.229	**0.000**	**0.016**	n/a	0.762
GFF	**0.000**	0.156	**0.042**	**0.000**	**0.001**	0.762	n/a

When we removed the algorithms in the statistical tests, the sample size was increased with reference to the number of algorithms. The algorithm plots in Fig. 6 illustrate that the relationships between the remaining algorithms still keep consistent. The relationships do not vary from the reduced number of algorithms in the experiments for case two study. This can also be observed from Tables 5–14. Although the corrected p-value changes, the relationships between the algorithms do not change. In other words, the image samples in case two are sufficient for the statistical tests.

Table 7
($1 \times N$) test (SWT and CVT removed) for the data in Table 2 (p-value$= 1.548761e-13$).

Fusion algorithm	LAP	NSCT	GRW	WSSM	HOSVD	GFF
Averaged ranking	3.3000*	3.8333*	5.7333*	4.2667*	2.2667	**1.6000**
Corrected p-value	0.01600758	0.001796474	7.216243e−09	0.0002369138	0.329114	NA



(a) SWT removed

(b) SWT and CVT removed

(c) SWT, CVT, and LAP removed

(d) SWT, CVT, LAP, and NSCT removed

(e) SWT, CVT, LAP, NSCT, and GRW removed

Fig. 6. Algorithm graphs for case study two by removing designated algorithms.

Table 8
Corrected p-values from ($N \times N$) pair-wise comparison (SWT and CVT removed) with algorithm ranks in case two.

	LAP	NSCT	GRW	WSSM	HOSVD	GFF
LAP	n/a	0.658	**0.003**	0.628	0.261	0.051
NSCT	0.658	n/a	**0.032**	0.658	0.087	**0.008**
GRW	**0.003**	**0.032**	n/a	0.095	**0.000**	**0.000**
WSSM	0.628	0.658	0.095	n/a	**0.020**	**0.001**
HOSVD	0.261	0.087	**0.000**	**0.020**	n/a	0.658
GFF	0.051	**0.008**	**0.000**	**0.001**	0.658	n/a

Table 9
($1 \times N$) test (SWT, CVT, and LAP removed) for the data in Table 2 (p-value$= 3.361554e-07$).

Fusion algorithm	NSCT	GRW	WSSM	HOSVD	GFF
Averaged ranking	3.1333*	4.8000*	3.5333*	1.9333	**1.6000**
Corrected p-value	0.01053512	1.192307e−07	0.001623575	0.5637029	NA

4.1.3. Case study three

The third case study is from reference [37], where the proposed fusion algorithm is using dictionary-based sparse representation (DSR). Three fusion metrics, Q_{abf}, *VIFF*, and *NMI* were employed in the study (see Appendices for the details). Five fusion algorithms

Table 10
Corrected p-values from ($N \times N$) pair-wise comparison (SWT, CVT, and LAP removed) with algorithm ranks in case two.

	NSCT	GRW	WSSM	HOSVD	GFF
NSCT	n/a	**0.016**	0.977	0.056	**0.032**
GRW	**0.016**	n/a	0.056	**0.000**	**0.000**
WSSM	0.977	0.056	n/a	**0.017**	**0.005**
HOSVD	0.056	**0.000**	**0.017**	n/a	0.977
GFF	**0.032**	**0.000**	**0.005**	0.977	n/a

Table 11
($1 \times N$) test (SWT, CVT, LAP, and NSCT removed) for the data in Table 2 (p-value$= 1.212787e-06$).

Fusion algorithm	GRW	WSSM	HOSVD	GFF
Averaged ranking	3.8667*	2.8667*	1.6667	**1.6000**
Corrected p-value	4.565973e−06	0.01079484	0.8875371	NA

Table 12
Corrected p-values from ($N \times N$) pair-wise comparison (SWT, CVT, LAP, and NSCT removed) with algorithm ranks in case two.

	GRW	WSSM	HOSVD	GFF
GRW	n/a	0.068	**0.000**	**0.000**
WSSM	0.068	n/a	**0.022**	**0.022**
HOSVD	**0.000**	**0.022**	n/a	0.888
GFF	**0.000**	**0.022**	0.888	n/a

Table 13
(1 × N) test (SWT, CVT, LAP, NSCT, and GRW removed) for the data in Table 2 (p-value= 0.01135255).

Fusion algorithm	WSSM	HOSVD	GFF
Averaged ranking	2.7333*	1.6667	**1.6000**
Corrected p-value	0.003817899	0.8551321	NA

Table 14
Corrected p-values from (N × N) pair-wise comparison (SWT, CVT, LAP, NSCT, and GRW removed) with algorithm ranks in case two.

	WSSM	HOSVD	GFF
WSSM	n/a	**0.006**	**0.006**
HOSVD	**0.006**	n/a	0.855
GFF	**0.006**	0.855	n/a

Table 15
Quantitative assessments of different multi-focus image fusion methods (modified from Table.1 in [37]). DSR is the proposed method.

Image	Fusion Metric	DCHWT	WSSM	CBF	IFM	GFF	DSR
1	Q_{abf}	0.6939	0.7094	0.7258	0.7328	0.7328	0.7395
1	*VIFF*	0.9026	0.9421	0.9234	0.9391	0.9391	0.9443
1	*NMI*	0.9771	1.1542	1.0969	1.1123	1.1123	1.2501
2	Q_{abf}	0.6549	0.6659	0.6991	0.7245	0.7256	0.7391
2	*VIFF*	0.8301	0.8695	0.8639	0.8781	0.8838	0.8811
2	*NMI*	0.8361	0.8737	0.9203	1.0943	0.9731	1.1503
3	Q_{abf}	0.6625	0.6708	0.7121	0.7384	0.738	0.748
3	*VIFF*	0.8531	0.9005	0.8899	0.9112	0.915	0.9176
3	*NMI*	1.0025	1.0241	1.069	1.2221	1.1332	1.2684
4	Q_{abf}	0.6876	0.7311	0.7205	0.7459	0.7511	0.762
4	*VIFF*	0.8018	0.9128	0.8562	0.9135	0.923	0.9269
4	*NMI*	0.6288	0.7733	0.7261	0.9054	0.8135	1.0091
5	Q_{abf}	0.7527	0.7289	0.7689	0.771	0.7777	0.782
5	*VIFF*	0.812	0.8853	0.8413	0.8788	0.8806	0.8738
5	*NMI*	0.9657	1.0291	1.0188	1.1056	1.0393	1.2522
6	Q_{abf}	0.5692	0.6092	0.6159	0.6333	0.6557	0.7087
6	*VIFF*	0.7157	0.7382	0.7546	0.6979	0.7657	0.7766
6	*NMI*	0.5296	0.5566	0.5825	0.7275	0.6179	0.9471

were selected for comparison. Table 15 shows the original assessment results.

The organize of the experimental data is quite similar to the data in study case two. For this case, we did a general (1 × N) and (N × N) tests, if significant differences exist, and will not get into the details for each image set or metric as we did for case two. The Friedman rank sum test (with Iman Davenport extension) obtained a p-value $2.2e-16$, which indicates the differences existing in the fusion algorithms. Fig. 7 plots the correlation among the fusion methods.

Table 16 gives the (1 × N) test results. Again, the star (*) illustrates the significant difference between the proposed DSR algorithms and the rests, which is supported by the corrected p-values in the table. The “best” averaged ranking is highlighted with bold font. The (N × N) test results are presented in Table 17 and Fig. 8. The difference between (1 × N) and (N × N) tests is observed. The (1 × N) test confirmed the significant difference between the proposed fusion algorithm and the other methods while the (N × N) test showed the connection between DSR and GFF. According to [32], the (1 × N) result is preferred as it requires less measurement points than (N × N) procedures in order to provide meaningful output. In this case, the difference between the proposed algorithm and GFF algorithm is not significant. An interesting chain results as the proposed fusion algorithm is not significant to GFF, GFF is not significant to IFM, IFM is not significant to WSSM, and WSSM is not significant to CBF or DCHWT.

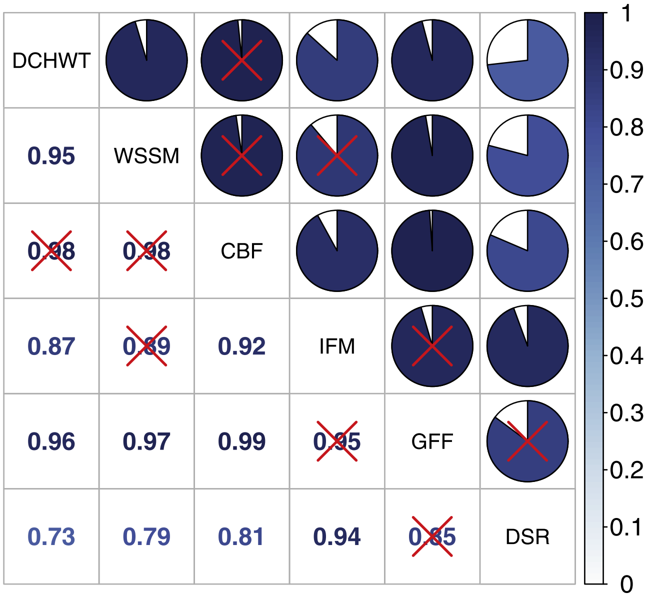


Fig. 7. The plot of the correlation matrix of the algorithms in case three. The cross indicates the correlation is not significant.

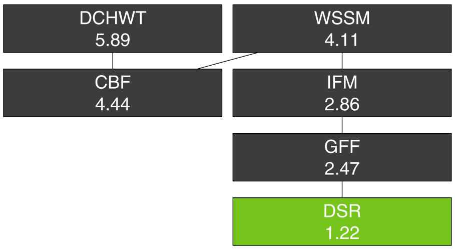


Fig. 8. Algorithm graph for all pair-wise comparison in case three.

4.2. Discussions

The first case study demonstrates the use of Wilcoxon signed rank test for comparison of two specific algorithms. However, most research may consider more than two fusion algorithms. Thus, the Wilcoxon signed rank test may not be adopted very often. When a multiple algorithm comparison is planned, three factors need to be considered, i.e. the data size, number of algorithms, and choice of metrics. In the second case, although the statistical tests were conducted for specific image group or fusion metric, the test results are not consistent due to a number of reasons. The fusion metrics are different and thus the assessments of the fusion performance may vary. When all the eight algorithms are considered for a specific image or metric, there are no sufficient data to obtain a reliable test result. Therefore, a general (1 × N) or (N × N) comparison is recommended. This is applicable to case three as well.

In image fusion research, (1 × N) comparison is recommended when proposing a new fusion algorithm. The new algorithm serves as a control or reference in the comparison. The (N × N) comparison can be applied to parameter tuning for fusion algorithm optimization. The results with different fusion parameter settings can be differentiated with this comparison.

For the significant tests, there is no established agreement about the minimum number of samples. The Wilconxon's test is less influenced by the sample size than Friedman's test [10,41]. A large sample size is preferred since the power of the statistical tests will increase. As a thumb of rule, the sample size, which doubles the number of algorithms, should be sufficient according to [10,41]. Similarly, more images will benefit the power of the

Table 16
(1 × N) test (Friedman test + Finner's correction) for the data in Table 15.

Fusion algorithm	DCHWT	WSSM	CBF	IFM	GFF	DSR
Averaged ranking	5.8889*	4.1111*	4.4444*	2.8611*	2.4722*	**1.2222**
Corrected p-value	3.619327e−13	6.020445e−06	5.945419e−07	0.01072234	0.04502088	n/a

Table 17
Corrected p-values from (N × N) pair-wise comparison with algorithm ranks in case three.

	DCHWT	WSSM	CBF	IFM	GFF	DSR
DCHWT	n/a	**0.026**	0.062	**0.000**	**0.000**	**0.000**
WSSM	**0.026**	n/a	1.000	0.090	**0.034**	**0.000**
CBF	0.062	1.000	n/a	**0.044**	**0.009**	**0.000**
IFM	**0.000**	0.090	**0.044**	n/a	1.000	**0.034**
GFF	**0.000**	**0.034**	**0.009**	1.000	n/a	0.090
DSR	**0.000**	**0.000**	**0.000**	**0.034**	0.090	n/a

statistical tests for the application of image fusion assessment. For the case study one, as only two algorithms were considered in the comparison, the Wilconxon signed rank test was carried out on nine image samples. For the case study two, five fusion metrics were calculated from three image databases. Thus, for the eight fusion algorithms, there are totally fifteen "samples" available for the tests (see Table 2). From the experiments in the subsection (4.1.2), we can see the statistical test results are consistent when we increased the ratio of the sample size to the number of fusion algorithms by removing fusion algorithms sequentially from the list. Therefore, the samples are sufficient for this case study when all the eight algorithms are considered in the tests, where the sample to algorithm ratio is about 1.9. In case study three, there are eighteen samples available for the tests of six algorithms. However, we do not have a quantitative number for the power of the statistical tests against sample size. As a general conclusion, the more image samples the better for the statistical test.

5. Conclusion

This paper proposes applying statistical analysis to the performance assessment in the image fusion research. The general procedure of the proposed statistical comparison consists of two steps: 1) detect the difference among fusion algorithms with a statistical significant test, and 2) identify the source of the differences with a corresponding *post-hoc* analysis. Three case studies chosen from recent publications are carried out with the proposed analytic method, which helps to identify the significance of a proposed algorithm. However, the algorithm comparison is a scientific problem as the result may vary with a number of factors, such as algorithms to compare as well as the metrics used for it. Moreover, the sample size may also affect the conclusion. Based on the general considerations for a non-parametric test [9], a larger sample size is preferred so that the power of the statistical test can be achieved. Statistical comparison is recommended to the image fusion research community, which can avoid duplicated research by assembling varied elements in the fusion framework.

Our future work will consider a more comprehensive study, in which an extensive statistical comparison of the state-of-the-art fusion algorithms will be carried out through screening the new fusion algorithms in recent publications. Furthermore, the Image fusion Toolbox Employing Significance Testing (ImTEST) made available through this research would be available for the community to assist in image fusion performance assessment.

Appendix

Tables 18 and 19 list the abbreviations for image fusion algorithms and fusion metrics respectively.

Table 18
Fusion algorithms in experiments.

Abbreviation	Full term	Reference
MAV	maximum of absolute values	[42]
SD	standard deviation	[14]
SWT	stationary wavelet transform	[43]
CVT	curvelet transform	[44]
LAP	Laplacian pyramid	[45]
NSCT	nonsubsampled contourlet transform	[46]
GRW	generalized random walks	[47]
WSSM	wavelet-based statistical sharpness measure	[48]
HOSVD	high order singular value decomposition	[15]
GFF	guided filtering based fusion	[49]
DCHWT	discrete cosine harmonic wavelet transform	[50]
CBF	cross bilateral filter	[51]
IFM	image matting	[52]
DSR	dictionary-based sparse representation	[37]

Table 19
Fusion metrics in the experiments.

Abbreviation	Full term	Reference
MI	mutual information	[53,54]
ENL	equivalent number of looks	[53,54]
CC	correlation coefficient	[53,54]
Q_y	Yang's metric	[55]
Q_c	Cvejie's metric	[56]
Q_g	gradient-based fusion metric	[57]
Q_p	Zhao's metric	[58]
Q_{mi} /NMI	normalized mutual information	[59]
Q_{abf}	gradient based quality index	[60]
VIFF	visual information fidelity for fusion	[61]

References

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

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

[3] Z. Liu, E. Blasch, Multisensor Data Fusion: From Algorithm and Architecture Design to Applications, Taylor & Francis, pp. 453–488.

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

[5] A. Toet, M. Hogervorst, S. Nikolov, J. Lewis, T. Dixon, D. Bull, C. Canagarajah, Towards cognitive image fusion, Inf. Fusion 11 (2) (2010) 95–113.

[6] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganière, W. Wu, Objective assessment of multiresolution fusion algorithms for context enhancement in night vision: a comparative study, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 94–109.

[7] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[8] S. Garcia, A. Fernandez, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power, Inf. Sci. 180 (10) (2010) 2044–2064. Special Issue on Intelligent Distributed Information Systems

[9] J. Derrac, S. Garcia, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm Evol. Comput. 1 (1) (2011) 3–18.

[10] S. García, A. Fernández, J. Luengo, F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft comput. 13 (10) (2008) 959–977.

[11] J. Luengo, S. Garcia, 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 Syst. Appl. 36 (4) (2009) 7798–7808.

[12] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2014) 147–164.

[13] K. Sharma, M. Sharma, Image fusion based on image decomposition using self-fractional fourier functions, Signal Image Video Process. 8 (7) (2014) 1335–1344.

[14] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.

[15] J. Liang, Y. He, D. Liu, X. Zeng, Image fusion using higher order singular value decomposition, Image Process. IEEE Trans. 21 (5) (2012) 2898–2909.

[16] X. Li, S.Y. Qin, Efficient fusion for infrared and visible images based on compressive sensing principle, IET Image Process. 5 (2) (2011) 141–147.

[17] H. Hariharan, A. Koschan, B. Abidi, A. Gribok, M. Abidi, Fusion of visible and infrared images using empirical mode decomposition to improve face recognition, in: Proceedings of International Conference on Image Processing, 2006, pp. 2049–2052.

[18] S. Li, J.T. Kwok, Y. Wang, Multifocus image fusion using artificial neural networks, Pattern Recognit. Lett. 23 (8) (2002) 985–997.

[19] Y. Xia, M. Kamel, Novel cooperative neural fusion algorithms for image restoration and image fusion, Image Process. IEEE Trans. 16 (2) (2007) 367–381.

[20] A. Toet, J. Walraven, New false color mapping for image fusion, Opt. Eng. 35 (3) (1996) 650–658.

[21] S. Yin, L. Cao, Y. Ling, G. Jin, One color contrast enhanced infrared and visible image fusion method, Infrared Phys. Technol. 53 (2) (2010) 146–150.

[22] Z. Xue, R.S. Blum, Concealed weapon detection using color image fusion, in: Proceedings of 6th International Conference of Information Fusion, 1, 2003, pp. 622–627.

[23] Y. Zheng, W. Dong, E.P. Blasch, Qualitative and quantitative comparisons of multispectral night vision colorization techniques, Opt. Eng. 51 (8) (2012) 087004–1–087004–16.

[24] Y. Zhang, Adaptive region-based image fusion using energy evaluation model for fusion decision, Signal Image Video Process. 1 (3) (2007) 215–223.

[25] T. Chan, S. Esedoglu, F. Park, A. Yip, Recent developments in total variation image restoration, in: N. Paragios, Y. Chen, O.D. Faugeras (Eds.), Handbook of Mathematical Models in Computer Vision, Springer-Verlag US, 2006, pp. 17–31.

[26] V. Caselles, Total variation based image denoising and restoration, in: Proceedings of the International Congress of Mathematicians, volume 3, Madrid, Spain, 2006, pp. 1453–1472.

[27] W.-W. Wang, P.-L. Shui, X.-C. Feng, Variational models for fusion and denoising of multifocus images, Signal Process. Lett., IEEE 15 (2008) 65–68.

[28] D.M. Greig, B.T. Porteous, A.H. Seheult, Exact maximum a posteriori estimation for binary images, J. R. Stat. Soc.. Series B (Methodol.) 51 (2) (1989) 271–279.

[29] M. Kumar, S. Dass, A total variation-based algorithm for pixel-level image fusion, IEEE Trans. Image Process. 18 (9) (2009) 2137–2143.

[30] C.H. Seng, A. Bouzerdoum, M. Amin, S.L. Phung, Two-stage fuzzy fusion with applications to through-the-wall radar imaging, Geosci. Remote Sensing Lett., IEEE 10 (4) (2013) 687–691.

[31] P. Dalgaard, Introductory Statistics with R, Statistics and Computing, 2, Springer, New York, USA, 2008.

[32] B. Trawinski, M. Smetek, Z. Telec, T. Lasota, Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms, Int. J. Appl. Math. Comput. Sci. 22 (4) (2012) 867–881.

[33] R.L. Iman, J.M. Davenport, Approximations of the critical region of the Friedman statistic, Commun. Stat. - Theory Methods 9 (6) (1980) 571–595.

[34] D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 4, Chapman & Hall / CRC, 2007.

[35] B. Calvo, G. Santafe, Scmamp: statistical comparison of multiple algorithms in multiple problems, R J. 8 (1) (2016) 1–8.

[36] X. Li, S.Y. Qin, Efficient fusion for infrared and visible images based on compressive sensing principle, IET Image Process. 5 (2) (2011) 141–147.

[37] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84.

[38] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, Keel: a software tool to assess evolutionary algorithms for data mining problems, Soft comput. 13 (3) (2008) 307–318.

[39] Z. Liu, E. Blasch, Statistical analysis of the performance assessment results for pixel-level image fusion, in: 17th International Conference on Information Fusion, Salamanca, Spain, 2014, pp. 1–8.

[40] I. Rodriguez-Fdez, A. Canosa, M. Mucientes, A. Bugarín, STAC: a web platform for the comparison of algorithms using statistical tests, in: 2015 IEEE International Conference on Fuzzy Systems, 2015, pp. 1–8.

[41] S. García, D. Molina, M. Lozano, F. Herrera, A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: aâ case study on the CEC'2005 special session on real parameter optimization, J. Heuristics 15 (6) (2008) 617–644.

[42] T. Wan, Z. Qin, An application of compressive sensing for image fusion, in: Proceedings of the ACM International Conference on Image and Video Retrieval, in: CIVR '10, ACM, New York, NY, USA, 2010, pp. 3–9.

[43] O. Rockinger, Image sequence fusion using a shift-invariant wavelet transform, in: Proceedings of International Conference on Image Processing, Vol. 3, 1997, pp. 288–301.

[44] L. Tessens, A. Ledda, A. Pizurica, W. Philips, Extending the depth of field in microscopy through curvelet-based frequency-adaptive image fusion, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 1, 2007, pp. I–861–I–864.

[45] P. Burt, E. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.

[46] Q. Zhang, B. long Guo, Multifocus image fusion using the nonsubsampled contourlet transform, Signal Process. 89 (7) (2009) 1334–1346.

[47] R. Shen, I. Cheng, J. Shi, A. Basu, Generalized random walks for fusion of multi-exposure images, IEEE Trans. Image Process. 20 (12) (2011) 3634–3646, doi:10.1109/TIP.2011.2150235.

[48] J. Tian, L. Chen, Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure, Signal Process. 92 (9) (2012) 2137–2146.

[49] S. Li, Z. Yao, W. Yi, Frame fundamental high-resolution image fusion from inhomogeneous measurements, Image Process. IEEE Trans. 21 (9) (2012) 4002–4015.

[50] B.K. Shreyamsha Kumar, Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform, Signal Image Video Process. 7 (6) (2013) 1125–1143, doi:10.1007/s11760-012-0361-x.

[51] B.K. Shreyamsha Kumar, Image fusion based on pixel significance using cross bilateral filter, Signal Image Video Process. 9 (5) (2015) 1193–1204, doi:10.1007/s11760-013-0556-9.

[52] S. Li, X. Kang, J. Hu, B. Yang, Image matting for fusion of multi-focus images in dynamic scenes, Inf. Fusion 14 (2) (2013) 147–162.

[53] N. Cvejic, C.N. Canagarajah, D.R. Bull, Image fusion metric based on mutual information and tsallis entropy, Electron. Lett. 42 (11) (2006).

[54] T.D. Dixon, E.F. Canga, J.M. Noyes, T. Troscianko, S.G. Nikolov, D.R. Bull, C.N. Canagarajah, Methods for the assessment of fused images, ACM Trans. Appl. Percept. 3 (3) (2006) 309–332.

[55] C. Yang, J. Zhang, X. Wang, X. Liu, A novel similarity based quality metric for image fusion, Inf. Fusion 9 (2008) 156–160.

[56] N. Cvejic, A. Loza, D. Bul, N. Canagarajah, A similarity metric for assessment of image fusion algorithms, Int. J. Signal Process. 2 (3) (2005) 178–182.

[57] C.S. Xydeas, V. Petrovic, Objective pixel-level image fusion performance measure, in: Proceedings of SPIE, 4051, 2000, pp. 89–98.

[58] J. Zhao, R. Laganiere, Z. Liu, Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement, Int. J. Innovative Comput. Inf. Control 3 (6(A)) (2007) 1433–1447.

[59] M. Hossny, S. Nahavandi, D. Vreighton, Comments on 'information measure for performance of image fusion', Electron. Lett. 44 (18) (2008).

[60] C.S. Xydeas, V. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.

[61] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2) (2013) 127–135.