Pakistan Journal of Science (Vol. 68 No. 1 March, 2016)

COMPARING MULTIPLE CORNER DETECTION ALGORITHMS USING NON-PARAMETRIC STATISTICAL TESTS

F. Anjum, N. Kanwal, Z. Altaf and A. Shaukat

Department of Computer Science, LCWU, Lahore, Pakistan
Corresponding Author email: fatimaanjum04@gmail.com

ABSTRACT: Performance characterization of algorithms has been commonly performed using parametric methods such as Precision-Recall, Repeatability, and detection rate *etc.* These methods assume that the data to be normally distributed, and therefore, the results became data specific. The main objective of this analytical study was to employ non-parametric statistical tests, for this purpose two tests Wilcoxon Signed Rank test and McNemar's test were applied to characterize the performance of corner detection algorithms. The results showed that the use of sufficiently large amount of data and correct testing framework using different non-parametric statistical tests yielded similar results, which was not observed with conventional parametric tests. Both Wilcoxon Signed Rank test and McNemar's test produced a similar ranking of corner detection algorithms, as both tests suggested Harris and Stephens at the top position then SUSAN, FAST, GLC and finally KLT. Hence, these non-parametric test were recommended to be used for the evaluation of vision algorithm due to their simplicity and reliability.

Keywords: Wilcoxon Signed Ranked test, McNemar's test, pair-wise comparison, corner detection, Performance evaluation.

(Received 19-10-2015 Accepted 12-03-2016)

INTRODUCTION

Considering the domains of machine learning, computer vision and image processing, the most common performance evaluation methods are Receiver Operating Characteristic (ROC) curves (Simonyan *et al.*, 2014), Precision-Recall curves (Miksik and Mikolajczyk, 2012), accuracy plots, repeatability graphs (Heinly *et al.*, 2012) and some statistical methods such as Analysis of Variance (ANOVA), t-tests *etc*. (Winer *et al.*, 1971). There are two main problems using these measures. First, although these performance measures are able to highlight an algorithm's success or failure, but one at a time, depending on the evaluation criterion used. Perhaps this is the reason that a variety of algorithms' ranking is found in the literature. Furthermore, graphs cannot be interpreted properly as in case of cross curve and overlapping curves so the reliability and statistical significance becomes questionable (Lobo *et al.*, 2008).

Secondly, the measures themselves are not flawless. A systematic analysis of a number of performance measures was performed by (Sokolova *et al.*, 2009). As per their analysis most of these performance measures are sensitive to the amount of data divided into positive and negative examples. Therefore, the results should be interpreted differently with change in data. Hence, testing method as well as framework of collecting algorithms' results should be as much non-specific as possible.

Statistical tests are categorized as parametric and non-parametric methods; distinction between two comes from the data, where former methods commonly assumes that the data to be normally distributed (Winer *et al.*, 1971) while the latter one ignores the data characteristics. Moreover, it is difficult to apply parametric tests to real data because data is generally non-normally distributed and need transformation. On the other hand non-parametric methods do not make any pre-assumption and more convenient to be used. In particular, they may be applied in situations where data characteristics are less known.

Although both parametric and non-parametric tests are used for multiple comparisons, however, the comparison of an algorithm with at least one existing state-of-the-art algorithm is typically performed known as paired or 1 x 1 comparison (Durkalski *et al.*, 2003). There are a number of statistical tests available for pair-wise comparisons such as t-test, Pearson test, Wilcoxon sign test and McNemar's test *etc*. The first two are parametric while the last two are non-parametric form of tests (Gibbons and Chakraborti, 2011). Therefore, this study employed McNemar and Wilcoxon tests for the performance assessment of corner detection algorithms. McNemar's test has been widely used in medical research (Durkalski *et al.*, 2003, Gonen *et al.*, 2001, Saag *et al.*, 1992, Uemura *et al.*, 2001 and Wellner *et al.*, 2004); however, it has not been commonly examined for characterizing vision algorithms.

MATERIALS AND METHODS

A problem of corners detection in digital images had been chosen for comparing corner detection algorithms and McNemar’s and Wilcoxon Signed Rank test for paired data analysis. Furthermore, to investigate the effect of the amount of data over evaluation results, the two selected tests were more suitable, because the Wilcoxon test could differentiate performance differences using a small amount of data while McNemar’s test needed large sample size as shown in Figure-1.

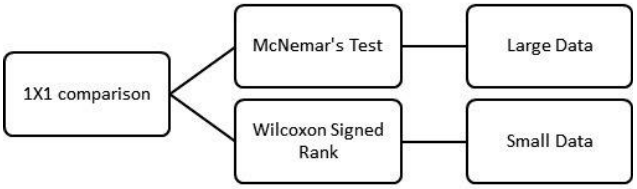

Figure 1. Non-parametric tests for paired data analysis

Image Data: In order to use sufficiently large amount of data both synthetic and real image data had been used for the evaluation of corner detectors that was originally developed to assess the angular sensitivity of corner detectors (Kanwal *et al.*; 2011a and Kanwal *et al.*, 2011b). But here the focus of analysis was not to identify a detector which could find corner points at all angles rather an identification of algorithm which could classify all image pixels appropriately. Collecting large amount of data was the primary consideration to obtain statistically reliable results. Therefore, geometric shapes were used, the purpose of using geometric shapes was that these shapes were simple and fundamental for the representation of shapes/objects in digital images as is shown in Figure-2. Also geometric information such as angle remained same while changing the orientation of a geometric object.

A large number of synthetic images of geometric shapes were digitally produced on computer. Polygons and stars like geometric shapes were generated on a computer and printed to generate real image data. Some of these images shown in Figure-2 are photographed using a Nikon D300 camera.

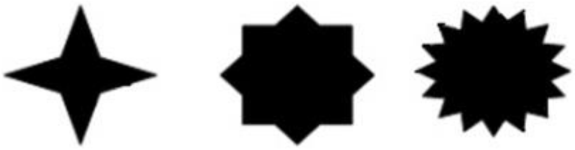

Figure 2. Different polygons used to generate synthetic and real image data.

To find out exact corner locations in photographed images, a two step procedure was followed. First 10 humans were asked to point out the pixel locations of corner points in photographed images. These locations were then refined by OpenCv’s function to find sub pixel accuracy. A total of 45 real and 45 synthetic images were used for this purpose.

For evaluation purpose, instead of finding only corner pixel locations, all pixels in an image were defined to be either corner or non-corner and stored as validation image as is shown in Figure-3.

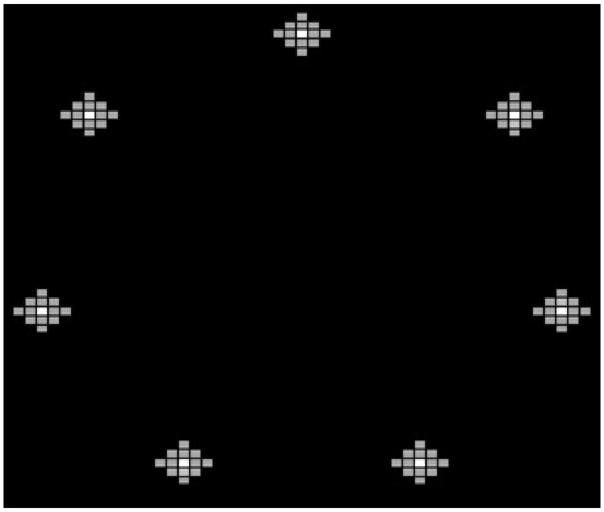

Figure 3: An image of a polygon with marked corner and its neighbourhing pixels with white and gray pixels respectively

The actual corner locations were stored as a white pixel value of 255, while black pixels valued 0, were non-corner pixels and gray color pixels valued 200 were the neighborhood pixels. This kind of validation images were generated for all synthetic and real images. In order to access individual detector’s performances, the detectors’ outcome was compared using these synthetic and real validation images.

To explore the effectiveness of two non-parametric statistical tests, a combination of classical and newly proposed corner detection algorithms were used which included Harris and Stephens (Harris and Stephens, 1988) and Kanade-Lucas-Tomasi (KLT) (Tomasi and Kanade, 1991) Smallest Univalue Segment Assimilating Nucleus (SUSAN) (Smith and Brady, 1997) whereas, features from Accelerated Segment Test (FAST) (Rosten *et al.*, 2010) and Global and Local Curvature Points (GLC) were two newly proposed corner detectors (He and Yung, 2008).

Testing Framework: Statistical analysis started by setting a null hypothesis denoted by H_o, A null hypothesis is a statement that describes the relationship between dependent and independent variable *e.g.* “the algorithms under study behaved similar to some selected data”. This hypothesis was then tested using different statistical tests and either accepted or rejected. For all statistical tests the results were interpreted using a level of significance called alpha (α). It was used to set a threshold or a cut off value for significant and non-significant results. For instance test statistics were translated as follows:

If a Z-score was less than Z-critical, the result was considered non-significant and the null hypothesis was accepted. However, if the Z-score was greater than or equal to Z-critical then the result was considered statistically significant and the stated null hypothesis was rejected for a given alpha; for which commonly used value was 0.05 *i.e.* 95% confidence limit.

Table 1. Truth table for McNemar's test

	Algo. A Failed	**Algo. A Passed**
Algo. B Failed	*ff*	*Ft*
Algo. B Passed	*tf*	*Tt*

f=false, *t*=true

Mcnemar's Test: The test was used to record the outcomes of two algorithms over multiple tests and therefore not only it counted the number of times an algorithm was successful or unsuccessful in a test but also the total number of tests performed.

For example, in comparing two algorithms, *i.e.* Algorithm A to Algorithm B, a null hypothesis was based on the presumption that there was no difference between their performances. To count pass and fail cases for truth table (Table 1), following method was adopted: if a corner detector detects a corner pixel in an original image, it was matched against three values in the validation image. If the detected location has value 0 in validation image, the detector failed and validation image passed. However, if the detected location had value 255 or 200, both detector and validation image passed. For the rest of the pixels, both detector and validation images were considered a pass as the detector did not detect these background pixels as the corner pixels. The Z score was calculated using following formula

$$\mathrm{z} = \sqrt{\chi 2} => \frac{(|\mathrm{tf} - \mathrm{ft}| - 1)}{\sqrt{\mathrm{tf} + \mathrm{ft}}}$$

and interpreted as: Z ≈0 showed that both Algorithms gave similar results. However, when Z increase the expression may involved cases where the output of one algorithm was true and the other was false.

Wilcoxon Signed Ranked Test: Similar to McNemar's test the Wilcoxon Signed Ranked test was also used for the comparison of two related samples. This test was used to test a Null Hypothesis stated as below

H_o: the median of the sample was zero

H_1: the median response of the sample was less than the median

Following were the steps to apply wilcoxon signed ranked test

1. Let n be the sample size for N number of pairs. For i = 1,…, N, let x_{1i} and x_{2i} represents the measurements or observations.
2. For i = 1,..., n, calculated difference between observations for each data set, i.e. $|x_{2i}-x_{1i}|$ and $\mathrm{sgn}(x_{2i}-x_{1i})$, where sgn represented sign function.
3. Eliminate pairs with $|x_{2i}-x_{1i}| = 0$. Let n_r be the decreased sample size.
4. Let W+ represents Sum of the ranks of positive differences, and W- represents sum of the ranks of negative differences. Check that $W_+ + W_- = 1/2*k(k+1)$, where k was the number in the sample having ignored the zeros.
5. For two tailed test pick the smallest of W+ and W-.
6. For one tailed test W- was the test statistics.
7. Calculate Z using formula $\frac{(\mathrm{W}-0.5-\mathrm{n_r}*(\mathrm{n_r}+1)/4)}{\sqrt{\mathrm{n_r}*(\mathrm{n_r}+1)*(2*\mathrm{n_r}+1)/24}}$ for $n_r > 15$.
8. Z greater than Z critical (taken from Z-table for given α) showed a significant performance difference between Pair of algorithms and corresponding W score pointed out the one performing better than the other.

RESULTS AND DISCUSSION

The classification of image pixels by the detectors was counted, which yielded binary results. Both McNemar's and Wilcoxon test used this count of binary outcomes. In both tests two algorithms were compared at a time, *i.e*. 1 x 1 comparisons, therefore, there was only one degree of freedom. Hence, from standard Z-tables the critical Z-score was 1.96 for α = 0.05.

First row of each detector correspond to Z value and second row always indicate corresponding P value. The Z score greater than 1.96 and P less than σ, for σ = 0.05 highlighted significant performance difference between two algorithms.

Mcnemar's Test Results: For convenience and quick overview, detectors' results for both synthetic and real images are shown in Table 2 and the arrowheads pointing towards the detector performing better in pair-wise comparison (at the intersection of row and column). Counting the arrowheads pointing in the same direction helped identifying Harris and Stephens to be the best algorithm followed by SUSAN similar to (Mokhtarian and Mohanna, 2006). FAST appeared to be next in the ranking order while KLT showed the worst performance as compared to all other detectors.

Wilcoxon Test Results: Similar to McNemar's test, Wilcoxon signed rank test was also used to compare algorithms in pairs. Furthermore, one tailed prediction was used to not only identify the performance difference, but also the one with better performance. Therefore, W-was the statistics that was used and the directions of arrowheads indicated better performance as is shown in Table-3. For all comparisons, Wilcoxon test appeared to be in complete agreement with McNemar's test, (Harris and Stephens, 1988) being the best of all algorithms. Moreover, both Wilcoxon and McNemar's test pointed out the better performing algorithm in pair-wise comparisons (Gibbons and Chakraborti, 2011).

Table 2. McNemar's Test results for synthetic and real images.

	Susan		Fast		GLC		KLT	
	Synthetic	**Real**	**Synthetic**	**Real**	**Synthetic**	**Real**	**Synthetic**	**Real**
	←7.038	1.510	←18.74	0.74	←2.645	←17.11	←15.86	←203.72
Harris	1.96E-12	1.31E-01	0.00	0.46	8.17E-03	0.00	0.00	0.00
Susan			←17.52	←2.215	←2.608	←18.673	←16.450	←2.04E+02
			0.00	2.67E-02	9.11E-03	0.00	0.00	0.00
Fast					↑8.043	↑19.805	←7.214	←203.72
					8.88E-16	0.00E+00	5.46E-13	0.00
GLC							←13.82	←196.12
							0.00	0.00

Table 3. Wilcoxon Test Results for synthetic and real image data.

		Susan		Fast 12		GLC		KLT	
		Synthetic	**Real**	**Synthetic**	**Real**	**Synthetic**	**Real**	**Synthetic**	**Real**
Harris	**z=**	←3.703	0.562	←4.95	0.021	←3.709	←5.750	←4.546	←5.838
	p=	2.13E-04	5.74E-01	7.50E-07	9.84E-01	2.08E-04	8.93E-09	5.47E-06	5.28E-09
	w+	716.50	467.50	847.50	497.00	717.00	988.00	920.50	1035.00
	w-	144.50	567.50	55.50	493.00	144.00	2.00	114.50	0.00
Susan			**z=**	←5.138	←2.73	←3.936	←5.750	←4.892	←5.838
			p=	2.79E-07	6.25E-03	8.29E-05	8.93E-09	9.96E-07	5.28E-09
			w+	935.50	533.00	734.50	988.00	914.50	1035.00
			w-	54.50	170.00	126.50	2.00	75.50	0.00
Fast 12					**z=**	↑4.201	↑5.84	←2.189	←5.838
					p=	2.65E-05	5.28E-09	2.86E-02	5.29E-09
					w+	106.00	1035.00	291.50	0.00
					w-	755.00	0.00	654.50	1035.00
GLC							**z=**	←4.350	←5.838
							p=	1.36E-05	5.29E-09
							w+	94.50	0.00
							w-	766.50	1035.00

Z-critical = 1.96 for α = 0:05 when applied for one tailed prediction. The arrowheads therefore, points the best performing algorithm and have larger W- value.

As discussed before, Wilcoxon Test was designed for a small amount of data. Therefore, similarity in McNemar's and Wilcoxon's test results supported the use these tests with more confidence. Moreover, if the researcher had sufficiently large data, then he/she could confidently use McNemar's test as it was much simpler and easily applicable than Wilcoxon test (Kanwal *et al.*, 2011b).

Similarly, the general testing framework helped inferring general ranking of corner detection algorithms which was different from the ones presented by (Tomasi and Kanade, 1991; Smith and Brady, 1997; Rosten *et al.*, 2010 and He and Yung, 2008). Due to the application specific criteria used in these studies such as the number of true corners detected by (Smith and Brady, 1997), repeatability of the corner points by (Rosten *et al.*, 2010) and localization accuracy by (He and Yung, 2008). Scores produced by both tests are presented in Table 4 and 5 were used to generate the ranking of corner detection algorithms is shown in Table- 6. Algorithms with maximum number of pointing arrowheads secured highest position. Which appeared to be able to differentiate between corner and non-corner pixels followed by SUSAN (Harris and Stephens, 1988). KLT algorithm showed a significant performance difference from all other algorithms for synthetic and real image data. The close position of FAST and SUSAN showed that circular mask based methods were more effective identify pixels even in the presence of noise as compared to curvature scale space algorithm.

Table 6. Ranking of Corner Detection algorithms based on Statistical Comparisons

1st	Harris and Stephens
2nd	SUSAN
3rd	FAST
4th	GLC
5th	KLT

Moreover, both Wilcoxon and McNemar's (Gibbons and Chakraborti, 2011) test pointed out the better performing algorithm in pair-wise comparison for which these could be used as post-hoc procedures for multiple comparison (1 x N) tests such as Friedman test and Quade tests reported by (Theodorsson, 1987). The agreement of these two tests on multiple tests results not only proved the reliability of these non-parametric statistical methods but also highlighted the easy application of Wilcoxon and McNemar's test for comparing vision related algorithms as well as in medical research as has been reported by (Durkalski *et al.*, 2003, Gonen *et al.*, 2001, Saag *et al.*, 1992, Uemura *et al.*, 2001 and Wellner *et al.*, 2004) where these tests were most commonly used.

Conclusions: In this work, the use of non-parametric statistical tests was encouraged for analyzing different algorithms due to the fact that the preconditions that guaranteed the reliability of the parametric tests were not satisfied with real data. Similarly, using the correct testing framework for identifying an algorithm with the best performance, pair-wise tests were found to be suitable and equally reliable. Comparison of different corner detection algorithms yielded a ranking, showing Harris and Stephens to be performing better than all other algorithms even the newly developed ones followed by SUSAN. Furthermore, the use of non-application based testing framework helped to generate a general ranking of algorithms instead of application of specific performance assessment.

REFERENCES

Durkalski, V. L., Y. Y. Palesch, S. R. Lipsitz and P. F. Rust. (2003). Analysis of clustered matched-pair data. Statistics in medicine, 22(15): 2417–2428.

Gibbons, J. D., and S. Chakraborti, (2011). Nonparametric statistical inference., pp. 977-979, Springer Berlin Heidelberg.

Gönen, M., K.S. Panageas and S.M. Larson. (2001). Statisticalissuesin analysis of diagnostic imaging experiments with multiple observations per patient1. Radiology, 221(3):763–767.

Harris, C. and M. Stephens. (1988). A combined corner and edge detector. In Alvey vision conference, Vol. 15, p. 50, Manchester, UK.

He, X. and N., Yung. (2008). Corner detector based on global and local curvature properties. Optical Engineering, 47:057008.

Heinly, J., E. Dunn, and J. M. Frahm. (2012). Comparative evaluation of binary features. In Computer Vision–ECCV (pp. 759-773). Springer Berlin Heidelberg.

Kanwal, N., S. Ehsan, E. Bostanci, and A. F. Clark. (2011a). Evaluating the angular sensitivity of corner detectors. In IEEE International Conference on Virtual Environments Human-Computer Interfaces and Measurement Systems (VECIMS), pages 1–4.

Kanwal, N., S. Ehsan, E. Bostanci, and A. F. Clark. (2011b). A statistical approach for comparing the performances of corner detectors. In IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim), pages 321– 326.

Lobo, J., A. Jiménez-Valverde, and R. Real. (2008). AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 17(2):145–151.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2):153–157.

Miksik, O. and K. Mikolajczyk. (2012). Evaluation of local detectors and descriptors for fast feature matching. In 21st International Conference on Pattern Recognition (ICPR), (pp. 2681-2684).

Mokhtarian, F. and F. Mohanna. (2006). Performance evaluation of corner detectors using consistency and accuracy measures. Computer Vision and Image Understanding, 102(1):81–94.

Rosten, E., and T. Drummond (2006). Machine learning for high-speed corner detection. In Computer Vision–ECCV (pp. 430-443), Springer Berlin Heidelberg.

Rosten, E., R. Porter, and T. Drummond. (2010). FASTER and better: A machine learning approach to corner detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 32:105–119.

Saag, M., W. Powderly, G. Cloud, P. Robinson, M. Grieco, P. Sharkey, S. Thompson, A. Sugar, C. Tuazon, J. Fisher. (1992). Comparison of amphotericin B with fluconazole in the treatment of acute AIDS-associated cryptococcal meningitis. New England Journal of Medicine, 326(2):83–89.

Simonyan, K., A. Vedaldi and A. Zisserman. (2014). Learning local feature descriptors using convex optimisation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(8), 1573-1585.

Smith, S. and J. Brady. (1997). SUSAN: A new approach to low level image processing. International Journal of Computer Vision, 23(1):45–78.

Sokolova, M. and G. Lapalme. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4):427–437.

Theodorsson-Norheim, E. (1987). Friedman and Quade tests: BASIC computer program to perform nonparametric two-way analysis of variance and multiple comparisons on ranks of several related samples. Computers in biology and Medicine, 17(2), 85-99.

Tomasi, C. and T. Kanade. (1991). Detection and Tracking of Point Features. Image Rochester NY, 91-132.

Uemura, N., S. Okamot. S. Yamamoto, N. Matsumura, S. Yamaguchi, M. Yamakido, K. Taniyama, N. Sasaki, and R. Schlemper. (2001). Helicobacter pylori infection and the development of gastric cancer. New England Journal of Medicine, 345(11):784.

Wellner, B., A. McCallum, F. Peng, and M. Hay. (2004). An integrated, conditional model of information extraction and coreference with application to citation matching. In Proceedings of the 20th conference on Uncertainty in artificial intelligence, pages 593–601. AUAI Press.

Winer, B. J., D. R. Brown, and K. M. Michels. (1971). Statistical principles in experimental design, volume 2. McGraw-Hill New York. Pages. 132-150.