Final Report of GREC'11 Arc Segmentation Contest: Performance Evaluation on Multi-resolution Scanned Documents

Hasan S.M. Al-Khaffaf[1,2], Abdullah Zawawi Talib[2], and Mohd Azam Osman[2]

[1] Image Understanding and Pattern Recognition Research Group (IUPR), Technische Universität Kaiserslautern, Kaiserslautern, Germany
[2] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Malaysia
hasansm@rhrk.uni-kl.de, {azht,azam}@cs.usm.my

Abstract. This paper presents the final report of the outcome of the sixth edition of the Arc Segmentation Contest. The theme of this edition is segmentation of images with different scanning resolutions. The contest was held offline before the workshop. Nine document images were scanned with three resolutions each and the ground truth images were created manually. Four participants have provided the output of their research prototypes. Two prototypes are more established while the other two are still in development. In general, vectorization methods produces better results with low resolution scanned images. Participants' comments on the behavior of their methods are also included in this report. A website devoted for this edition of the contest to hold the newly created dataset and other materials related to the contest is also available.

Keywords: Empirical Performance Evaluation, Graphics Recognition, Raster to Vector Conversion, Line Drawings, Statistical Analysis.

1 Introduction

This edition of the Arc Segmentation Contest 2011 was held in conjunction with the Ninth IAPR International Workshop on Graphics RECognition (GREC) held in Chung-Ang University, Seoul, Korea, in September 2011. This contest was organized by the School of Computer Sciences, Universiti Sains Malaysia, Malaysia. The theme of this contest is multi scanning resolutions and its effect on line detection. The test images were selected from a text book. Three scanning resolutions were employed: 200, 300, and 400 DPI. Ground truth data were created manually using a vector editor and they were stored in the VEC file format. The output of research prototypes as well as commercial software were accepted. Two methods of participation were available. The first option involves using the DAE platform [1,2] while the second option was participation through email. In both cases the output of participating methods should be in the VEC or DXF file formats.

Y.-B. Kwon and J.-M. Ogier (Eds.): GREC, LNCS 7423, pp. 187–197, 2013.

2 Test Images, Ground Truthing, and Expected Vectors

A set of mechanical engineering drawings were selected from an old textbook [3]. The selected images were then scanned by AstraSlim scanner with three different resolutions: 200, 300, and 400 DPI. In the scanning process, trial and error were employed to ensure that the graphical elements in the paper drawing were captured with minimum rotation angle. The scanned color images were then cropped and binarized with 50% threshold. We have nine images for each scan resolution and a total of 27 test images. The images are shown in Fig. 1.

For the generation of the ground truth data, we started with images with high resolution (400 DPI). Vectors were created for each corresponding raster entity (arc or circle) in the image using a vector editor. Contextual knowledge was put into effect during the creation of the vector data such as arcs/circles that are co-centered and an arc/circle passing through the center of another arc/circle. After creating the vectors for all necessary primitives, the vectors were combined into one block and copied to the image with lower resolutions (300 and 200 DPI). The block of vector entities were then fitted manually on the raster image by continuous resizing/moving till it fitted well on the corresponding raster image. We opted to retain the text strings in the scanned image. The reason for not whiting out the text strings is to make the vectorization scenario as close as possible to real situations. Additionally, all text strings in the original raster images were not vectorized. This dataset does not support line width information, hence the width for all ground truth arcs/circles is set to the value of 1.

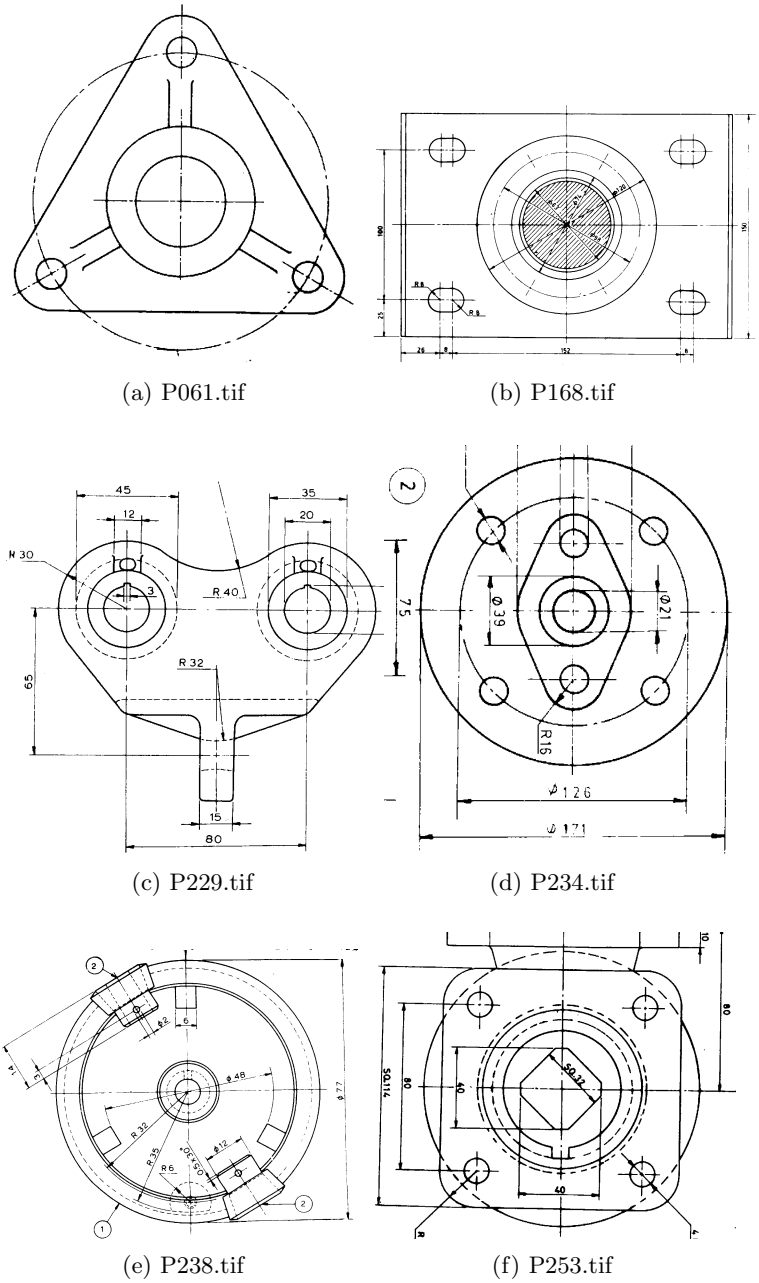
As in the previous editions of the arc segmentation contest, the focus is on circle and arc detections only. For this reason, straight lines were ignored. The text strings were also ignored and not saved to the vector file. Dashed circle/arc entities are not supported. However, circular segments that are parts of dashed circles/arcs and large enough to be recognized can be saved as separate arcs. Each of these arcs will has center, radius, start angle, and end angle. The detected circles were stored as circle entities identified by a center and a radius in the output vector file. The pixel is the unit of measure and the top-left corner of the image/screen is at point (0,0).

A website[1] is available for the contest. The test images generated in this contest were hosted in the website as well as within the DAE platform (Section 3).

3 Methods of Participation

As with previous edition of the contest [4], research prototypes as well as commercial software were accepted in this edition of the contest. However, all the four participants (as shown in Table 1) have provided the output of their own research prototypes. Nevertheless, commercial software were also tested on the newly created dataset [5]. Two options of participation were available:

[1] `http://www.cs.usm.my/arcseg2011/`

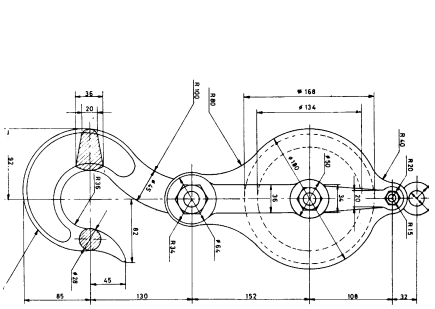


(a) P061.tif (b) P168.tif

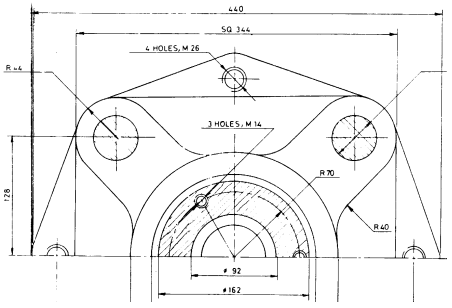
(c) P229.tif (d) P234.tif

(e) P238.tif (f) P253.tif

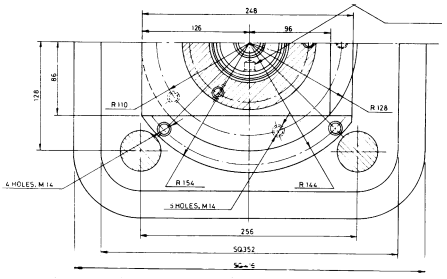
Fig. 1. Test images (Scanned from Sidheswar et al. [3])



(g) P254.tif



(h) P260A.tif



(i) P260B.tif

Fig. 1. (*Continued)*

1. The participants were expected to provide the output of their systems through the DAE platform[2] of Lehigh University.
2. The participants were expected to provide us (through e-mail) with the output of the vectorization methods in the VEC or DXF formats.

The participants can adopt either of the two options based on their preferences. Because of simplicity of using the second option and some technical problems, the second option was the preferred option in this contest. Finally, the output of Liu's method is included for comparison purpose only and not considered officially participating in the contest since the performance evaluation measure are also from the same author. This way we will remove any bias towards other contestants.

4 Performance Evaluation Method

In this edition we continue to use VRI [8] as the performance evaluation index. VRI (in the range [0..1]) is calculated as follows:

$$VRI = \sqrt{D_v * (1 - F_v)} \tag{1}$$

[2] DAE platform [1,2]: `http://dae.cse.lehigh.edu/dae/`

Table 1. Participated Methods

Vectorizer	Author(s)	Affiliation
Liu's method [6]	Liu Wenyin	Department of Computer Science, City University of Hong Kong, Hong Kong, China
ArcFind† 3.1	Dave Elliman	School of Computer Science, University of Nottingham, UK
Effective Arc Segmentation† (EAS) 1.0	Zili Zhang, Xuan Wang, Yanjun Ma	Shenzhen Graduate School, Harbin Institute of Technology, ShenZhen, China
Qgar-Lamiroy [7]	Bart Lamiroy	Université de Lorraine, LORIA, Nancy, France

†Unpublished work

where D_v is the detection rate and F_v is the false alarm rate. A high VRI score indicates better recognition rate. Under this measure, the detected vector quality is the geometric mean of five factors: endpoints quality, overlap distance quality, line width quality, line style quality, and line shape quality. In this contest the dataset does not include width information hence the line width quality factor is eliminated (neutralized) when calculating the VRI index.

5 Results and Discussion

For each of the participated methods, we obtained 27 VRI scores (nine images * three resolutions) and a total of 108 scores for the four participated methods. Table 2 shows performance scores of the four participated methods. In terms of stability, the two best performers, Liu and Qgar-Lamiroy are stable with one exception for the latter where it fails (VRI=0) at image P229-400dpi. These two methods are more mature than the other two and have participated in past editions of this contest. For the other two methods ArcFind and EAS, the former is more stable (fails on one image P229-400dpi) while the latter fails on five images. These two methods are currently under development[3] which could be one of the reasons for getting this sort of performance. As mentioned in §3, Liu's method is included for comparison purposes only. Objectively and in terms of VRI scores, the Qgar-Lamiroy method is the highest performer, and hence it is the winner of this contest.

In order not to limit ourselves to finding a winner for the contest, we opted to perform a more rigorous analysis on the effect of scanning resolution on the performance of the participated methods. In this edition of the contest, we have studied the resolution factor as well as the vectorization factor. Instead of performing a superficial test on the data and relying on the mathematical mean, we opted to go a step further and used a robust statistical analysis. The use of statistical test will also provide answers to research questions that were not

[3] The argument is based on email communication with the authors.

Table 2. Performance scores [D_v, F_v, VRI] for the participated methods

Image	Liu's	ArcFind	EAS	Qgar-Lamiroy
P061†	[.770, .174, **.798**‡]	[.144, .845, .149]	[.240, .599, .310]	[.286, .160, .490]
	[.690, .331, **.679**]	[.288, .924, .148]	[.226, .525, .328]	[.243, .315, .408]
	[.616, .292, **.660**]	[.436, .909, .200]	[0, 1, 0]	[.347, .186, .532]
P168	[.705, .227, **.738**]	[.429, .869, .237]	[.034, .807, .081]	[.072, .775, .127]
	[.470, .541, **.464**]	[.222, .936, .119]	[.050, .894, .073]	[.320, .547, .381]
	[.285, .695, .295]	[.126, .973, .058]	[.016, .880, .044]	[.306, .623, **.340**]
P229	[.553, .422, **.565**]	[.494, .800, .314]	[.196, .693, .245]	[.354, .225, .524]
	[.429, .547, .441]	[.352, .895, .192]	[.238, .650, .288]	[.384, .239, **.540**]
	[.327, .681, **.323**]	[0, 1, 0]	[.015, .979, .018]	[0, 1, 0]
P234	[.290, .486, **.386**]	[.145, .950, .085]	[.178, .339, .343]	[.121, .545, .235]
	[.658, .257, .699]	[.176, .937, .105]	[.201, .716, .239]	[.695, .100, **.791**]
	[.469, .496, .486]	[.236, .912, .144]	[.150, .314, .321]	[.687, .133, **.772**]
P238	[.204, .718, .240]	[.221, .950, .105]	[.057, .750, .119]	[.139, .308, **.311**]
	[.397, .646, **.375**]	[.093, .974, .049]	[0, 1, 0]	[.146, .610, .239]
	[.284, .745, **.269**]	[.194, .970, .077]	[0, 1, 0]	[.139, .708, .201]
P253	[.909, .214, **.846**]	[.122, .966, .065]	[.130, .811, .157]	[.507, .355, .572]
	[.645, .399, **.623**]	[.136, .937, .093]	[.129, .587, .231]	[.507, .275, .606]
	[.488, .569, .459]	[.517, .855, .274]	[.060, .876, .086]	[.435, .419, **.503**]
P254	[.422, .406, **.501**]	[.327, .908, .174]	[.042 .643, .123]	[.106 .365, .260]
	[.368, .550, **.407**]	[.099, .970, .054]	[0, 1, 0]	[.178, .572, .276]
	[.525, .469, **.528**]	[.109, .978, .049]	[.034, .685, .103]	[.198, .594, .283]
P260A	[.467, .449, **.507**]	[.149, .959, .078]	[.037, .814, .083]	[.054, .213, .207]
	[.528, .471, **.528**]	[.396, .934, .162]	[.025, .903, .049]	[.205, .227, .398]
	[.525, .559, **.481**]	[.058, .989, .025]	[.087, .775, .140]	[.125, .576, .230]
P260B	[.379, .525, **.424**]	[.061, .976, .038]	[.097, .597, .198]	[.074, .498, .193]
	[.318, .647, **.335**]	[.114, .971, .057]	[0, 1, 0]	[.176, .700, .230]
	[.218, .769, **.224**]	[.024, .996, .009]	[.048, .618, .135]	[.084, .866, .106]
Avg	0.492	.113	.138	.361

†1st, 2nd, and 3rd rows of each image, corresponds to 200, 300, and 400 DPI
‡Highest VRI score in each resolution is shown in bold

possible to be answered using the mean or at least could not be answered with confidence. A precision of 95% is used for all the statistical tests of this paper.

In our experiment, we have two independent variables (*Method* and *Resolution*) and one dependent variable (*VRI*). Each paper image (subject) was used many times, and thus producing many VRI scores for any single paper drawing. Repeated measure ANOVA is the statistical test which is suitable to handle experiments with similar nature [9,10]. However, before starting with ANOVA, we need to check it's three requirements: (i) order effect should be avoided, (ii) the data should be normally distributed, and (iii) the Sphericity condition should not be violated.

The order effect is avoided since the original paper drawings were used in scanning and the raster images were not changed during the run of any software.

Shapiro-Wilk test [11] is used to check the normality condition. The null (H_0) and the alternative (H_1) hypotheses are as follows:

$$H_0 : \text{There is no difference between the distribution of the data and the normal.} \quad (2)$$

$$H_1 : \text{There is a difference between the distribution of the data and the normal.} \quad (3)$$

The ρ of the Shapiro-Wilk for all the cells (values are not shown in this paper) are not significant ($Sig. > .05$), indicating a failure to reject the null hypothesis, and hence the data are considered normally distributed.

In order to test the validity of the Sphericity condition, Mauchly's Test [12] needs to be performed. In Table 3, $\rho > .05$ for the two factors *Method* and *Resolution*; and the interaction between them, *Method* * *Resolution*. Hence the Sphericity condition is not violated.

Table 3. Mauchly's Test of Sphericity‡

Within Subjects Effect	df	Sig. (ρ)
Method	5	.408
Resolution	2	.280
Method * Resolution	20	.657

‡Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

After validating the three conditions for using repeated measure ANOVA, it will be possible to proceed with the statistical method by analyzing the Test of Within-Subjects Effects (Table 4). The *Method* variable is significant while the *Resolution* variable as well as the interaction between the two variables are not significant ($\rho \geq .05$). The significance of the *Method* variable means that we fail to reject the null hypothesis (H_0). On the other hand, the insignificance of the *Resolution* variable indicates that we reject the null hypothesis and accept the alternative one. The hypotheses for each of the variables are as follows:

$$H_0 : \mu_{\text{Liu}} = \mu_{\text{ArcFind}} = \mu_{\text{EAS}} = \mu_{\text{Qgar-Lamiroy}} \quad (4)$$

$$H_1 : \text{Not all the means are equal} \quad (5)$$

$$H_0 : \mu_{\text{200DPI}} = \mu_{\text{300DPI}} = \mu_{\text{400DPI}} \quad (6)$$

$$H_1 : \text{Not all the means are equal} \quad (7)$$

In other words, there are significant differences between the vectorization methods in terms of VRI scores. However, there are little differences between the

scanning resolutions in terms of VRI scores. It is shown in Table 5 that the mean differences between Liu method and the other three methods, Qgar-Lamiroy, EAS and ArcFind are significant ($\rho < .05$). Qgar-Lamiroy method is also significantly better than the other two methods EAS and ArcFind. However, Qgar-Lamiroy has a special case of high VRI scores with moderate resolution (300 DPI) while all the other methods get the best results with the lowest resolution (200 DPI). The performance of methods within the three resolutions is shown in Fig. 2.

With regard to scanning resolution, it is shown in Fig. 3 that the mean VRI scores drops with the increase in scanning resolution. However, the drop is not statistically significant.

Table 4. Tests of Within-Subjects Effects

Source	df	F	Sig. (ρ)
Method	3	54.33	.000
Error (Method)	24		
Resolution	2	2.39	.124
Error (Resolution)	16		
Method * Resolution	6	1.55	.183
Error (Method*Resolution)	48		

Sphericity Assumed

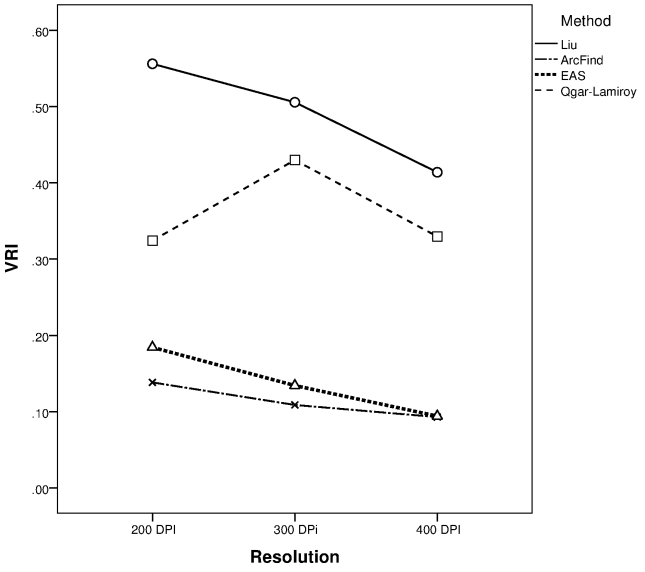


Fig. 2. Performance of vectorization methods with different scanning resolutions

Table 5. Pairwise Comparisons

(I) Method	(J) Method	Mean Difference (I-J)	Std. Error	Sig.[a]
Liu	ArcFind	.379	.035	.000
	EAS	.354	.038	.000
	Qgar-Lamiroy	.131	.035	.034
ArcFind	EAS	.024	.025	1.000
	Qgar-Lamiroy	-.248	.042	.002
EAS	Qgar-Lamiroy	-.224	.031	.001

Based on estimated marginal means

[a]. Adjustment for multiple comparisons: Bonferroni.

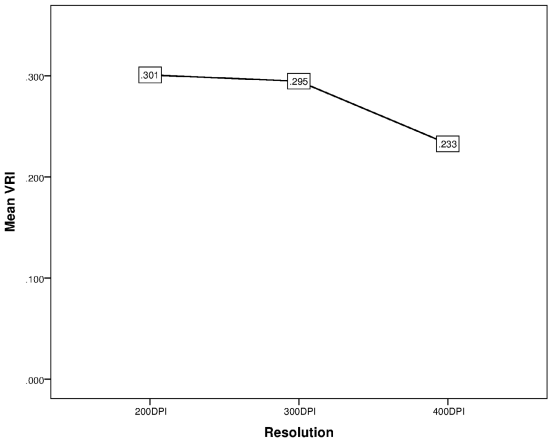


Fig. 3. VRI score means for the three resolutions

6 Post Contest Discussion

After the GREC event, we invited the authors to comment on the performance of their methods. In this section we highlight the outcome of the discussion. One paragraph is devoted to each author and their revised comments are presented next.

Qgar-Lamiroy method is based on matching an algebraic circle formula on a set of discrete points retrieved from the skeleton and measuring the overall fitting error. The higher the resolution, the more robust the skeleton is and the more precise the fitting error is measured. On the other hand, when the resolution gets higher, the algorithm starts detecting any small distortion in the shape of the circle (e.g. slightly oval in shape due to perturbations in the image processing chain whereby the human eye will not notice but the algorithm will). Although most parameters are scale/resolution invariant, there is one -tolerance on radius error- which is set in pixels, and therefore influencing the result when resolution becomes too high. Further work will eliminate these parameters (or at least try and make them scale invariant).

The ArcFind method is developed to detect circular arcs in scanned document images. It works well when arc shapes are close to full circle, but its performance drops when arcs are small and when short arcs are connected to straight lines causing small arcs to be detected as polyline. The contest images have a variety of arcs and circles that distract the ArcFind method and cause a drop in detection and hence causing it to get low VRI scores.

The EAS method has difficulty in obtaining accurate values for the center and radius of a circle. The method is better in detecting circles than in detecting arcs. The method also has difficulty in detecting arcs/circles that are tangents to other graphical elements.

Liu Wenyin commented on the experiment and gave feedbacks to improve its robustness. The feedbacks were incorporated in this report. However, no comments were provided on the performance of his method.

7 Summary

The outcome of the sixth edition of the Arc Segmentation Contest has been presented in this paper. The contest was held off-line before the GREC'11 workshop. Empirical performance evaluation of multi scan resolution was the theme of this contest. Four participants have provided the output of their own research prototypes. The highest performer in this contest is Qgar-Lamiroy method, and hence, it is the winner of this contest. One outcome of this contest is the creation of new multi-resolution ground truth data. This work is also the first study on research prototypes that involves multi resolution scanned images. The other outcome is in the new finding that increasing image resolution has negative effect on the performance of the tested methods. However, the drop in performance with higher scanning resolutions is not statistically significant. Actually, we have invited the authors to explain to us the reason behind any unusual performance of their methods. Lamiroy provides justification (See §6) on why his method gives good results with mid-resolution images. For two other authors, their methods are still under development and the details of their method are not published yet.

Acknowledgment. The authors appreciate the efforts of Abdul Halim Ghazali in preparing the ground truth images and Wong Poh Lee for creating and maintaining the contest website. We would also like to thank all participants for their contribution to the success of this contest. During the organization of the contest, the first author was a Post-Doctoral Fellow in the School of Computer Sciences, USM, Malaysia.

References

1. Lamiroy, B., Lopresti, D.: A Platform for Storing, Visualizing, and Interpreting Collections of Noisy Documents. In: Fourth Workshop on Analytics for Noisy Unstructured Text Data - AND 2010. ACM International Conference Proceeding Series. IAPR, ACM, Toronto, Canada (2010)

2. Lamiroy, B., Lopresti, D.: An Open Architecture for End-to-End Document Analysis Benchmarking. In: 11th International Conference on Document Analysis and Recognition - ICDAR 2011. International Association for Pattern Recognition, Beijing (2011)
3. Sidheswar, N., Kannaiah, P., Sastry, V.V.S.: Machine Drawing. Tata McGraw-Hill, New Delhi (1992)
4. Al-Khaffaf, H.S.M., Talib, A.Z., Osman, M.A., Wong, P.L.: GREC'09 Arc Segmentation Contest: Performance Evaluation on Old Documents. In: Ogier, J.-M., Liu, W., Lladós, J. (eds.) GREC 2009. LNCS, vol. 6020, pp. 251–259. Springer, Heidelberg (2010)
5. Al-Douri, B.A.T., Al-Khaffaf, H.S.M., Talib, A.Z.: Empirical Performance Evaluation of Raster to Vector Conversion with Different Scanning Resolutions. In: Badioze Zaman, H., Robinson, P., Petrou, M., Olivier, P., Shih, T.K., Velastin, S., Nyström, I. (eds.) IVIC 2011, Part I. LNCS, vol. 7066, pp. 176–182. Springer, Heidelberg (2011)
6. Liu, W.Y., Dori, D.: Incremental arc segmentation algorithm and its evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(4), 424–431 (1998)
7. Lamiroy, B., Guebbas, Y.: Robust and Precise Circular Arc Detection. In: Ogier, J.-M., Liu, W., Lladós, J. (eds.) GREC 2009. LNCS, vol. 6020, pp. 49–60. Springer, Heidelberg (2010)
8. Liu, W.Y., Dori, D.: A protocol for performance evaluation of line detection algorithms. Machine Vision and Applications 9(5-6), 240–250 (1997)
9. Roberts, M.J., Russo, R.: A Student's Guide to Analysis of Variance. Routledge (1999)
10. Al-Khaffaf, H.S.M., Talib, A.Z., Salam, R.A.: Empirical performance evaluation of raster-to-vector conversion methods: A study on multi-level interactions between different factors. IEICE Transactions on Information and Systems E94.D(6), 1278–1288 (2011)
11. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika 52(3/4), 591–611 (1965)
12. Mauchly, J.W.: Significance test for sphericity of a normal n-variate distribution. The Annals of Mathematical Statistics 11(2), 204–209 (1940)