INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT

IJRCM

A Monthly Double-Blind Peer Reviewed (Refereed/Juried) Open Access International e-Journal - Included in the International Serial Directories

Indexed & Listed at:

Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Cabell's Directories of Publishing Opportunities, U.S.A.,

Open J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)],

Index Copernicus Publishers Panel, Poland with IC Value of 5.09 & number of libraries all around the world.

Circulated all over the world & Google has verified that scholars of more than 3412 Cities in 173 countries/territories are visiting our journal on regular basis.

Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA

http://ijrcm.org.in/

CONTENTS

CALL FOR MANUSCRIPTS

We invite unpublished novel, original, empirical and high quality research work pertaining to recent developments & practices in the areas of Computer Science & Applications; Commerce; Business; Finance; Marketing; Human Resource Management; General Management; Banking; Economics; Tourism Administration & Management; Education; Law; Library & Information Science; Defence & Strategic Studies; Electronic Science; Corporate Governance; Industrial Relations; and emerging paradigms in allied subjects like Accounting; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Rural Economics; Co-operation; Demography: Development Planning; Development Studies; Applied Economics; Development Economics; Business Economics; Monetary Policy; Public Policy Economics; Real Estate; Regional Economics; Political Science; Continuing Education; Labour Welfare; Philosophy; Psychology; Sociology; Tax Accounting; Advertising & Promotion Management; Management Information Systems (MIS); Business Law; Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labour Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; International Relations; Human Rights & Duties; Public Administration; Population Studies; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism & Hospitality; Transportation Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic; Web Design and emerging paradigms in allied subjects.

Anybody can submit the **soft copy** of unpublished novel; original; empirical and high quality **research work/manuscript** ***anytime*** in ***M.S. Word format*** after preparing the same as per our **GUIDELINES FOR SUBMISSION**; at our email address i.e. infoijrcm@gmail.com or online by clicking the link **online submission** as given on our website (***FOR ONLINE SUBMISSION, CLICK HERE***).

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION:**

DATED: _____

THE EDITOR
IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF .**

(e.g. Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript entitled '_____' for possible publication in your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published elsewhere in any language fully or partly, nor is it under review for publication elsewhere.

I affirm that all the author (s) have seen and agreed to the submitted version of the manuscript and their inclusion of name (s) as co-author (s).

Also, if my/our manuscript is accepted, I/We agree to comply with the formalities as given on the website of the journal & you are free to publish our contribution in any of your journals.

NAME OF CORRESPONDING AUTHOR:
Designation:
Affiliation with full address, contact numbers & Pin Code:
Residential address with Pin Code:
Mobile Number (s):
Landline Number (s):
E-mail Address:
Alternate E-mail Address:

NOTES:

a) The whole manuscript is required to be in ***ONE MS WORD FILE*** only (pdf. version is liable to be rejected without any consideration), which will start from the covering letter, inside the manuscript.
b) The sender is required to mentionthe following in the **SUBJECT COLUMN** of the mail:
New Manuscript for Review in the area of (Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/ Engineering/Mathematics/other, please specify)
c) There is no need to give any text in the body of mail, except the cases where the author wishes to give any specific message w.r.t. to the manuscript.
d) The total size of the file containing the manuscript is required to be below **500 KB**.
e) Abstract alone will not be considered for review, and the author is required to submit the complete manuscript in the first instance.
f) The journal gives acknowledgement w.r.t. the receipt of every email and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of manuscript, within two days of submission, the corresponding author is required to demand for the same by sending separate mail to the journal.

2. **MANUSCRIPT TITLE:** The title of the paper should be in a 12 point Calibri Font. It should be bold typed, centered and fully capitalised.

3. **AUTHOR NAME (S) & AFFILIATIONS:** The author (s) **full name**, **designation**, **affiliation** (s), **address**, **mobile/landline numbers**, and **email/alternate email address** should be in italic & 11-point Calibri Font. It must be centered underneath the title.

4. **ABSTRACT:** Abstract should be in fully italicized text, not exceeding 250 words. The abstract must be informative and explain the background, aims, methods, results & conclusion in a single para. Abbreviations must be mentioned in full.

5. **KEYWORDS:** Abstract must be followed by a list of keywords, subject to the maximum of five. These should be arranged in alphabetic order separated by commas and full stops at the end.

6. **MANUSCRIPT:** Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. It must be prepared on a single space and single column with 1" margin set for top, bottom, left and right. It should be typed in 8 point Calibri Font with page numbers at the bottom and centre of every page. It should be free from grammatical, spelling and punctuation errors and must be thoroughly edited.

7. **HEADINGS:** All the headings should be in a 10 point Calibri Font. These must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

8. **SUB-HEADINGS:** All the sub-headings should be in a 8 point Calibri Font. These must be bold-faced, aligned left and fully capitalised.

9. **MAIN TEXT:** The main text should follow the following sequence:

INTRODUCTION

REVIEW OF LITERATURE

NEED/IMPORTANCE OF THE STUDY

STATEMENT OF THE PROBLEM

OBJECTIVES

HYPOTHESES

RESEARCH METHODOLOGY

RESULTS & DISCUSSION

FINDINGS

RECOMMENDATIONS/SUGGESTIONS

CONCLUSIONS

SCOPE FOR FURTHER RESEARCH

ACKNOWLEDGMENTS

REFERENCES

APPENDIX/ANNEXURE

It should be in a 8 point Calibri Font, single spaced and justified. The manuscript should preferably not exceed ***5000 WORDS***.

10. **FIGURES &TABLES:** These should be simple, crystal clear, centered, separately numbered &self explained, and **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. It should be ensured that the tables/figures are referred to from the main text.

11. **EQUATIONS:**These should be consecutively numbered in parentheses, horizontally centered with equation number placed at the right.

12. **REFERENCES:** The list of all references should be alphabetically arranged. The author (s) should mention only the actually utilised references in the preparation of manuscript and they are supposed to follow **Harvard Style of Referencing**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc, in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italics. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parentheses.
- The location of endnotes within the text should be indicated by superscript numbers.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–22 June.

UNPUBLISHED DISSERTATIONS AND THESES

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, Kurukshetra University, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITES

- Garg, Bhavet (2011): Towards a New Natural Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

SKEW DETECTION AND CORRECTION OF INDIAN SCRIPT DOCUMENTS

VIKRAM SHARMA
ASST. PROFESSOR
DAV COLLEGE
AMRITSAR

ABSTRACT

An optical character recognition (OCR) is the process of converting scanned images of machine printed documents into computer readable codes. Recognition of characters of Devnagari and Gurumukhi scripts have been an area of research for many peoples and large number of research papers and reports have been published in this area. The generalized way for developing OCR system of any script involves preprocessing, segmentation, feature extraction & classification and Post processing. To develop OCR for any language, the input document to be scanned may have some skew ness and needs to be tackled properly. This paper discuss in detail about skew angle detection and correction which is the main operation performed during Preprocessing phase of an OCR system. A fast and robust method for detecting and correcting skew angle is described which is based on head line which is inherent properties of Indian languages like Hindi and Punjabi. For estimation of skew angle, the proper headline of the digitized image is detected and its skew angle is estimated with horizontal direction. The proposed skew detection and correction method has accuracy about 89%.

KEYWORDS

Devnagari script, Preprocessing, Segmentation, Feature extraction, Classification, Post processing, skeletonizing, skew ness, Horizontal projection, Vertical projection.

INTRODUCTION

OCR is the acronym for Optical Character Recognition. A human eye can read the characters of any language written correctly by person or when it is printed. Making the computer do the same is called the problem of character recognition. OCR (optical character recognition) is the process of converting scanned images of machine printed or handwritten text or hand printed text into computer readable codes. Character recognition systems can contribute tremendously to the advancement of the automation process and can improve the interaction between man and machine in many applications including post offices, defense organizations, office automation, cheque verification and a large variety of banking, business, data entry applications and commercial data processing applications [1]. The main goal of OCR is to imitate the human ability to read in a very fast rate.

The process of recognizing characters of any script is usually performed through several phases which are as follows:

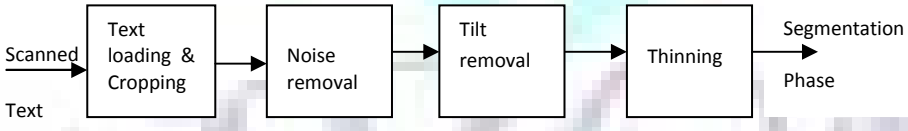
1. Pre-processing
2. Segmentation
3. Feature Extraction
4. Classification
5. Post processing

PREPROCESSING OF MACHINE PRINTED SCRIPT

In a typical OCR system, input document is read and digitized by an optical scanner. The Preprocessing phase of OCR takes digitized image as input, reduce noise and distortion, remove skew ness perform skeletonizing or thinning of the image and hence simplify the processing of remaining stages. The main advantage of preprocessing a text is to organize the information so as to make the task of character recognition simpler. Skew detection and correction of Indian script documents is a crucial part of preprocessing phase of OCR as it affects the subsequent understanding of the document.

The preprocessing starts with data acquisition in which documents to be processed are first scanned and digitized, and ends up with output to segmentation phase of character recognition system as shown in fig. 1. In this Figure, the various processes involved in preprocessing stage of OCR are shown. First the digitized text is loaded and cropped. Then the noise if any in the digitized text is removed. After this, skewness of the digitized text is corrected and then thinning is performed.

FIG. 1: FLOW DIAGRAM OF PRE-PROCESSING PHASE OF OCR



SKEW DETECTION AND CORRECTION

Skew ness in the digitized image is unavoidable because of casual use of scanner by human operator. Whenever a document is fed in to a scanner because of hand placement or automatic document feeder mechanisms, a few degree of skew (tilt) is always there. The skew angle is the angle that the text lines of the digitized image make with the horizontal direction. Hand placement or automatic document feeder mechanisms normally create $1 - 3^{\circ}$ of skew, due to the document's incorrect placement, or to a slight variation in roller speed. In some cases, the skew can reach as much as 10°. When the skew angle is $2 - 3^{\circ}$, the accuracy of document analysis and OCR is reduced; when it is more than 5, however, the result becomes unreliable. As a result, the skew estimation of document images to be carried out before segmentation and Feature extraction and classification phase of OCR [2, 3, 4].

Skew detection and correction are important for a successful OCR. Skew detection and correction can be achieved in two steps

(a) Estimation of skew angle
(b) Rotation of the image by skew angle in opposite direction.

Image of Gurumukhi document with some skew ness is shown in figure 2 and figure 3 shows the same document after correcting skew ness from it.

FIG. 2-SKEW NESS IN THE DOCUMENT

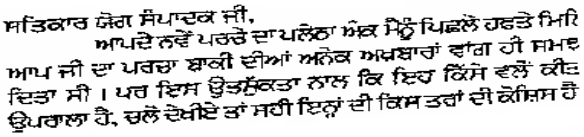
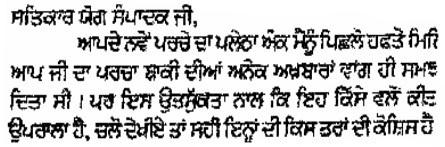


FIG. 3-SCRIPT WITHOUT SKEW NESS



REVIEW OF LITERATURE

In the past years many techniques have been applied to skew detection. One of the techniques is based on the projection profile of documents [5, 6]. For a document with horizontal text lines, the horizontal projection profile will have peaks at text line positions and troughs at positions in between successive text lines. To determine the skew of a document, the projection profile is computed at a number of angles and for each angle, a measure of difference of peak and trough height is made. The maximum difference corresponds to the best alignment with the text line direction which in turn, determines the skew angle. A nearest neighbor clustering based approach to estimate skew angle of the document is described by hashizume, yeh, Rasenfield [7]. In their methodology, all the connected components in the document were extracted and for each component, the direction angle of its nearest neighbor is estimated. A histogram of the direction angle is computed and the peak of which indicates the document skew angle. A heuristic approach has been applied by Lehal and Singh to detect and correct skew ness in Gurumukhi script [8].

PROPOSED METHOD FOR SKEW DETECTION AND CORRECTION

To detect and correct skew angle, a fast and robust method is based on head line which is inherent properties of Indian languages like Hindi and Punjabi. For estimation of skew angle, the proper headline of the digitized image is detected and its skew angle is estimated with horizontal direction. The detection of head line is necessary for script zone and character and word segmentation which will be used in later stages of OCR. So headline based skew detection and correction helps in saving computation during zone detection and character and word segmentation, which is main benefit of this approach. Head line is very useful for character segmentation from words and to divide a text line in different zones for the OCR of script like Devnagari and Bangla [9].

This method of detecting and correcting skew angle follows the following steps.

a) Check whether digitized document has skew ness or not
b) If yes then skew angle is detected with horizontal direction
c) Correct skewness

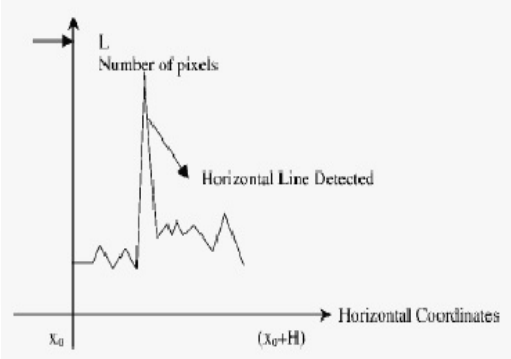
a) ***DETECTING DOCUMENT WITHOUT SKEW***

Scan the whole image horizontally and count the number of pixels in each row. Prepare a table with two columns of row number and number of pixels. We find out from the table whether some consecutive lines fulfill the constraint mentioned below. Based on this, we define a feature which once extracted will prove that the character image has a tilt or not. That feature has been named as 'line' which means that a feature 'line' has been detected when scanned row gives the number of information pixels that are at least 90% of the total number of pixels possible horizontally in an image frame as shown in figure 4.

H_{LINE} >= 0.9 * L (number of pixels)

where H_{LINE} is the number of pixels in a row while scanning horizontally.

FIG. 4: HORIZONTAL PROJECTION

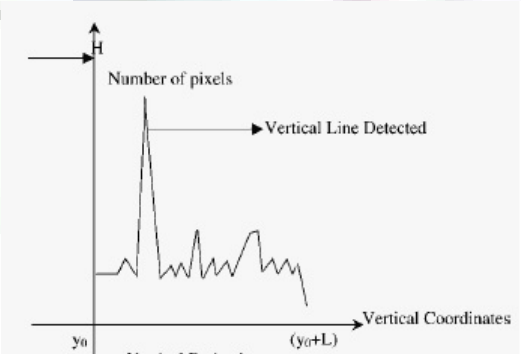


The X-axis in the graph shows the row number and the Y-axis shows the number of pixels. The graph showing the peak meeting of the constraint mentioned above will be termed as 'line detection' and once detected has to be confirmed through vertical scanning. Even if the line has been detected after horizontal scanning, it still has to be confirmed through vertical scan and no further processing is needed if the same result is obtained. Vertical projection and finding the line feature will also require the same effort as in the horizontal process. The same has been demonstrated in Figure 5.

V_{LINE} >= 0.9 * H (number of pixels)

where V_{LINE} is the number of pixels in a row while scanning vertically.

FIG. 5: VERTICAL PROJECTION



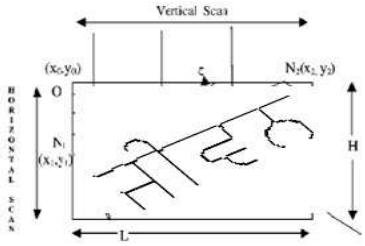
Once the 'line' has been detected, there is no further need to process the image to find the tilt. This means that the image is either exactly horizontally straight, i.e. at 0^o, 180^o or it is exactly vertical at 90^o or 270^o. The distinction between the 0^o and 180^o images will depend on the direction of the scan (top-left or top-right) and the detection of the line feature. Similar process is applied to have distinction between 90^o and 270^o. If the 'line' is not confirmed during the vertical scan then that means the image is tilted and we need to find the tilt angle and remove it.

b) SKEW ANGLE DETECTION

Characters in Hindi or Punjabi languages are formed in a structured manner. Whereas most of the language scripts are written with reference to the central axis, the Gurumukhi script is written w.r.t. the central axis along with an axis parallel to the central axis. This is the inherent property of writing Gurumukhi script. Once the angle of the axis w.r.t. the reference axis has been detected, thereafter the word can be rotated perfectly horizontally. This dominating feature of Indian script has been used for detecting the tilt angle. This can be better explained with an example (Fig. 6). Scannable area is nothing but the cover of the word

when it is viewed from four directions. But the informative scan area is the projected view ON1 from left side of the character image and ON2 is the projected view from topside of the image. If the word image is enclosed in a rectangular frame then, we easily get the length and the height of the image, which has been represented by L and H, respectively. Tilt angle has been represented by ζ. Figure 6 shows the skewed document being used to demonstrate the skew detection process.

FIG. 6: WORD SAMPLE



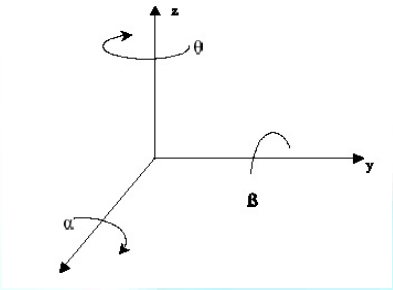
***c)* SKEW REMOVAL PROCESS AFTER SKEW DETECTION**

Once skew angle has been detected and finalized then a simple transformation technique works perfectly fine to straighten the tilted word of Devnagari or Gurumukhi script.

The transformation technique [4] mentioned below is applicable for rotating a point about the coordinate axes. To rotate a point about another arbitrary point in space requires three transformations. The first translates the arbitrary point to the origin, the second performs the rotation and the third translates the point back to its original position.

Referring to fig. 7, the rotation of a point about the *z* coordinate axis by an angle θ is achieved by using the transformation.

FIG. 7: TILT ANGLES



$$R_\theta = \begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R\alpha = \begin{bmatrix} 0 & \cos\alpha & \sin\alpha & 0 \\ 0 & -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where α= ζ

$$R\beta = \begin{bmatrix} \cos\beta & 0 & -\sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ \sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where β = ζ

CONCLUSION AND LIMITATION

A fast and robust method for detecting and correcting skew ness is described in this paper which is the main operation performed during preprocessing phase of character recognition. The method is based on head line which is inherent properties of Indian languages like Hindi and Punjabi. So headline based skew detection and correction helps in saving computation during zone detection and character and word segmentation, which is main benefit of this method. The proposed skew detection and correction method has accuracy about 89%. The limitation of this skew detection and correction method is that it will not work

well if the skew angle is very large. However, it has been observed that such skew angles are relatively uncommon. Another limitation is that this method of skew detection and correction does not provide desired results if the document is multi-skewed.

REFERENCES

1. B. B. Chaudhuri and U. Pal, "A complete printed devanagri OCR System"
2. F. N. Liolios and G. Kokkinakis, "On the generalization of the form identification and skew detection problem".
3. Y. Lu and C. L. Tan, "A nearest neighbor chain based approach to skew estimation in document images".
4. H. K. Kwag, S. H. Kim, S. H. Jeong and G. S. Lee, "Efficient skew estimation and correction algorithm for document images," Image and Vision Computing 20, 25–35
5. T. Akiyama and N. Hagita, "Automatic entry system for printed documents".
6. T. Pavlidis and J. Zhou, "Page segmentation and classification", Computer Vision Graphics and Image Processing.
7. Hashizume, P. S. Yeh and A. Rasenfeld, "A method of detecting the orientation of aligned components".
8. Lehal G. S. and Singh Chandan, OCR of Gurumukhi script.
9. A complete OCR system for continous Bengali characters by Raiman, Rahman.

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Computer Application & Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you tosupply your critical comments and suggestions about the material published in this issue as well as on the journal as a whole, on our E-mail**infoijrcm@gmail.com** for further improvements in the interest of research.

If youhave any queries please feel free to contact us on our E-mail infoijrcm@gmail.com.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

Sd/-
Co-ordinator

DISCLAIMER

The information and opinions presented in the Journal reflect the views of the authors and not of the Journal or its Editorial Board or the Publishers/Editors. Publication does not constitute endorsement by the journal. Neither the Journal nor its publishers/Editors/Editorial Board nor anyone else involved in creating, producing or delivering the journal or the materials contained therein, assumes any liability or responsibility for the accuracy, completeness, or usefulness of any information provided in the journal, nor shall they be liable for any direct, indirect, incidental, special, consequential or punitive damages arising out of the use of information/material contained in the journal. The journal, nor its publishers/Editors/ Editorial Board, nor any other party involved in the preparation of material contained in the journal represents or warrants that the information contained herein is in every respect accurate or complete, and they are not responsible for any errors or omissions or for the results obtained from the use of such material. Readers are encouraged to confirm the information contained herein with other sources. The responsibility of the contents and the opinions expressed in this journal is exclusively of the author (s) concerned.

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals