International Journal of Modern Engineering Research (IJMER)

www.ijmer.com Vol.1, Issue.2, pp-270-277 ISSN: 2249-6645

A Statistical Study on Optimal Usage of Intelligent Character Recognition Technology to Fetch Quality Data for Automation of University Exam System

Vikas Sharma

International Centre for Distance Education and Open Learning,
Himachal Pradesh University, Summer Hill, Shimla, India

ABSTRACT :

The automation of examination system for a typical university has been overwhelmed with various data quality problems because of involvement of manual data entry process from candidates' filled examination forms to evaluators' handwritten award lists, importing or exporting of data, limited resources, etc. Poor data quality can have a significant negative impact on organizations' success especially for a university where its entire credibility is dependent on accuracy and timely processing of results. As a result, organizations are implementing latest technologies to fetch quality data to achieve competitive advantage as well as to satisfy the varying needs of users. In this paper, a case study of Intelligent Character Recognition (ICR) system "AutoRec" is presented with regard to different data quality parameters such as accuracy, time, value added services, security, timeliness, etc. by processing a good sample of manual handwritten awards lists involving both numeric as well as alphanumeric characters. The results of the study indicate that the ICR based "AutoRec" system has the potential solution to improve data quality, minimize human intervention, reduce cost and time by balanced usage of scanning parameters, validation checks and confidence levels. Further, ICR is not the substitute of human operator but can minimize manual intervention.

Keywords - Data Quality, ICR, Scanning Parameters, Validation Checks, Confidence Levels, Human Intervention

1. INTRODUCTION

Information is increasingly becoming a critical asset for success in the modern societies throughout the world. Information is being created, processed, stored and retrieved and transmitted instantaneously from one end to another but the basic question is "how much this information is fit for use?" The data Quality (DQ) is one of the key determinants that decide the success or failure of any organisation. Errors in data cause variety of problems and raise costs in several areas. Earlier an error is detected, the cheaper it is to correct [1]. A typical university examination system consists of large data volumes, heterogeneous data types, widely distributed data sources and multiple stakeholders. The very existence of any university can be threatened by poor data quality (DQ). The data on which examination results are based and upon which the future of thousands of students depends if inaccurate, incomplete or has other types of problems can put a big question mark on the credibility. The automation of such examination system demands high data quality management system in place to avoid garbage-in-garbage out. The examination wing of Himachal Pradesh University, Summer Hill, Shimla compiles the results of the students by performing manual data entry of awards using a small software (utility) "Awards Management System" but this system doesn't provide quality data for processing. Further, the manual data entry is a slow, laborious, and expensive process compared to automatic recognition of text and its subsequent processing. Data entry specialists are becoming increasingly difficult to employ since it is tedious and boring work [5]. Computer has the ability to perform numerous tasks simultaneously and efficiently on scanned images by recognizing characters using artificial intelligence power. The most significant among these technologies is the Intelligent Character Recognition (ICR) for hand written documents and Optical Mark Recognition (OCR) for data capture from printed documents. The ICR is seemingly a good technology to fetch data from real world and convert into computer readable form. In this paper, a case study of newly introduced ICR system "AutoRec" viz.-a-viz. manual data entry system is conduced in the examination wing of Himachal Pradesh University to know which system can provide better quality of input data for processing by involving minimal human intervention, cost and time.

1. NEED OF INTELLIGENT CHARACTER RECOGNITION SYSTEM

Paper forms are still the least expensive data capture device where individuals without network connections must provide data for entry into a computer system. Even today's much-acclaimed Internet browser interfaces do not help computer applications when the data must be collected from

www.ijmer.com Vol.1, Issue.2, pp-270-277 ISSN: 2249-6645

constantly changing individuals and constantly changing locations [5]. The forms are easy to use even today due to little initial cost per form. No doubt that the documents are increasingly originated on the computer, however, in spite of this, it is unclear whether the computer has decreased or increased the amount of paper. Documents are still printed out for reading, dissemination, and markup [4]. Colleges and universities throughout the country are struggling to find some way to deal with paper documents that must be maintained to ensure institutional accountability. The improvement in hardware and increasing use of computers for storing paper documents has paved the way for document processing and recognition. The cost of optical scanners for document input have dropped to the level that these are affordable to even small businesses and individuals. In addition to above, the advancements in document analysis software and algorithms have also improved the text and image recognition rates significantly up to the level of 90 to 95% [4].

2. ISSUES RELATED TO DATA QUALITY AND COST USING INTELLIGENT CHARACTER RECOGNITION SYSTEM

Errors in data cause a variety of problems and raise the costs in several other associated areas. The cost to recognize and detect errors is not small whereas significant amount is involved to correct these data errors. The largest cost components are the hidden costs that affect the efficiency, productivity and public image of the organisation [1]. To maintain data quality has become the essential task for universities who have to compile huge volume of students data related with admission and examination processes. A small mistake in students' result status due to consideration of poor quality data can put the credibility of whole university under scanner which further also lowers the public image as well as involvement of litigation costs, etc. Since data quality is the major issue for data processing jobs so it become essential to design and tune character recognition applications to achieve high data quality. There are two types of recognition errors in ICR system: 1) rejected errors -unrecognized characters, and 2) substitution errors- erroneously recognized characters. Rejects need to be corrected by human intervention but substitutions must first be detected and then corrected [1]. Document scanners can misread an image that is dirty or too skewed. Characters read without contextual analysis may be interpreted as letters, when only numbers should exist in a field [5]. Substitution errors are the most dangerous because during these errors, an incorrect character is substituted for the correct character [3]. An image of the rejected character is presented to the data entry operator who corrects it by re-entering the actual character and program control automatically moves to the next rejected character. The rejected character cost is dependent on three factors: 1) missing characters- unreadable character, 2) extraneous characters, and 3) key entry speed of operator [1]. So, in nutshell, there are two factors which influence the cost to repair reject characters: 1) accuracy of the ICR recognition engine, and 2) reject re-entry speed. An increase in reject re-entry rate also leads to substantial error cost reduction. This is because the data entry operator can often key the entire field faster than repairing several rejects [1].

3. RESEARCH OBJECTIVES

- To study the opinions of technical staff about quality of input data essential for compilation of examination results.
- To analyse and compare quality dimensions of input data fetched through ICR based system "AutoRec" viz.-a-viz. manual data entry system "Awards Management System".
- To study the effect of image quality, confidence levels and validation checks on character recognition level of ICR based system "AutoRec".
- To analyse the effect of image quality, confidence levels and validation checks on cost, time and human intervention involved in ICR based system "AutoRec".

4. RESEARCH DESIGN AND METHODOLOGY

The research methodology of this study has been divided into three main parts, namely: 1) Scope of Study, 2) Population and Sample, and 3) Research Tools.

5.1 SCOPE OF STUDY

This study is conducted in the Examination Wing of Himachal Pradesh University and specifically on two examination classes - B.Com. and B.Sc. whose results are complied using computers.

5.2 POPULATION AND SAMPLE

To study the first two objectives, this study is based on convenient sample survey of nine technical staff members-four programmers and five data entry operators who actually use the ICR based "AutoRec" system- a product of FilFlan Technologies as well as in-house developed Awards Management System (AMS) based on Visual Basic 6.0 as front end and MS-Access 2000 as back end to fetch input data for compilation of results. To study the last two objectives, five samples of ICR compatible awards lists 2 from B. Com. part-III and 3 from B. Sc. Part -III, regular examinations, March 2010 were selected again using convenient sampling technique to know the effect of scanning parameters, validation checks and confidence levels on character recognition level of "AutoRec" system and involvement of human intervention, cost, time, etc. The

www.ijmer.com Vol.1, Issue.2, pp-270-277 ISSN: 2249-6645

above sample had 208 data fields which in turn summed to recognition of 761 characters. The opinion of ICR system "AutoRec" was classified using three point Likert Scale – matched characters, unrecognised characters and substituted characters.

5.3 RESEARCH TOOL

To study the first two objectives, the data collection tool was self designed questionnaire having two parts: 1) first part was used to rate the importance of input data quality parameters in context of University Examination System, and 2) second part was used to observe the quality dimensions of input data fetched through manual data entry system "AMS" as well as ICR based "AutoRec" system separately. A 5-point Likert Scale (5 = highly important and 1 = not important at all) was used to observe the opinion of the dealing persons corresponding to each quality dimension. To build an initial list of data qualities, the fifteen data quality dimensions defined [6] as: 1) access security, 2) accessibility, 3) accuracy, 4) appropriate amount of data, 5) believability, 6) completeness, 7) concise representation, 8) ease of understanding, 9) ease of understanding, 10) interpretability, 11) objectivity, 12) relevancy, 13) representational consistency, 14) reputation, and 15) timeliness were discussed in detail followed by brainstorming sessions to conclude a raw list relevant in context of compilation of results. These items were arranged then in logical order to give a questionnaire format. Using the literature on information/data quality and by looking carefully for overlap of data qualities in context of examination system, the items in the questionnaire were reduced to a more manageable 9 items with small description to provide readily available detail for observers while completing their questionnaires. The data collection process was carried out firstly by using the first part of the questionnaire followed by second part of questionnaire. Further to study the third objective, the following methods were used:

To analyse the effect of image quality on character recognition level of "AutoRec" system, four cases (T1C1, T1C2, T2C1 and T2C2) were designed using different scanning parameters where T1 (128 units), T2 (184 Units), C1 (128 units) and C2 (144 units) are default and best threshold 'T' and contrast- 'C' values on Fujitsu Scanner Fi4340C. The 128 units is the default value for threshold and contrast whereas for best visibility 184 units is the threshold value and for best sharpness the 144 units is the contrast value. Further, the red colour was dropped during scanning and the size of award lists used here was of legal (8.5"X14") size.

To study the effect of data validations checks on recognition accuracy of ICR system, two types of validation checks namely: 1) NVC (No Validation Checks), and 2) ONC (Only Numeric Checks) were applied on all above four cases (T1C1, T1C2, T2C1 and T2C2) separately.

- To study the recognition accuracy of ICR system at different confidence levels, four confidence levels (50, 75, 90 and 100 units) were experimented separately.
- To study the fourth objective, the response of ICR system based on three points Likert Scale (matched, unrecognised and substituted characters) was divided into two segments- characters needed human intervention (unrecognised and substituted characters) and characters needed no human intervention (matched characters). Total numbers of character needed human interventions were compared with actual number of characters which needed human intervention (manual data entry) to analyse the cost and time involved in both systems. In manual data entry system, double entry of every single award is done to have good accuracy level and to avoid any kind of discrepancy. The above observed opinions of the ICR system then converted into appropriate data tables and different statistical techniques were applied for analysis using MS-Excel 2007 spreadsheet.

5. RESULTS AND DISCUSSION

The table 1 shows the summary averages for weighted and unweighted data sets. Firstly, the importance score (IS) shows the average importance ranking for each question as rated by technical staff. Secondly, the average scores per data item for ICR based "AutoRec" system and manual data entry system (MDES) "Awards Management System (AMS)" are given. This is displayed in two modes: 1) raw score (RS) as unweighted ratings (with a theoretical Likert Scale range of 1 to 5), and 2) weighted score (WS). The weighted score (theoretically ranging from 1 to 25) is obtained by multiplying the unweighted score by the importance score for each respondent. The data quality considered most important by the technical staff, e.g. upper quartile (5.00) are all about accuracy, representational consistency and access security. The data qualities considered least important, e.g. below lower quartile (4.8) is value added features. Other quality dimensions are in between and the median for above importance score is 4.9.

The above total scores make it difficult to analyse the quality of input data fetched through above two system, so Data Quality Index (DQI) was calculated for each system using weighted score against the total possible score (Importance Score X 5). Overall, it appears that input data fetched through "AutoRec" system scored (0.9 points) followed by Manual Data Entry System (0.7 points). The Data Quality Index (DQI) for different input data quality dimensions is shown in fig. 1 for "AutoRec" system as well as Manual Data Entry System. Though there is no much

www.ijmer.com Vol.1, Issue.2, pp-270-277 ISSN: 2249-6645

more difference on some data quality dimensions such as appropriate amount of data, completeness, ease of understanding, representational consistency, value-added features and accessibility but a clear circle around by "AutoRec" system over Manual Data Entry System on other data quality dimensions such as accuracy, timeliness and access security indicates the difference.

TABLE I
Summary Averages for Weighted and Unweighted Data Sets

Sr. No.	Quality Dimensions	IS	SD	WIS	ICR "AutoRec"			MDES "AMS"		
					RS	WS	DQI	RS	WS	DQI
1	**Accuracy** Fetched Data is correct, i.e. free of errors.	5.0	0.0	25.0	4.2	21.1	0.84	2.9	14.5	0.58
2	**Appropriate Amount of Data** The quantity or volume of obtained data is appropriate for compilation of results	4.9	0.3	24.5	4.4	22.2	0.91	4.4	22.2	0.91
3	**Completeness** Fetched data is sufficient for compilation of results.	4.9	0.3	24.5	4.3	21.7	0.89	4.2	21.1	0.86
4	**Ease of Understanding** Obtained data is clear, without ambiguity and easy to comprehend.	4.8	0.4	23.9	4.6	22.8	0.95	4.4	22.2	0.93
5	**Timeliness** Time taken to convert manual data into digital form is reasonably good.	4.8	0.4	23.9	4.2	21.1	0.88	1.9	9.5	0.40
6	**Representational Consistency** Fetched data is represented in the specified format and compatible with previous data.	5.0	0.0	25.0	4.7	23.4	0.93	4.6	22.8	0.91
7	**Accessibility** Obtained data is available for usage easily and quickly.	4.9	0.3	24.5	4.1	20.6	0.84	3.6	17.8	0.73
8	**Value-Added** Obtained data is beneficial and provide advantages for value added services.	4.6	0.5	22.8	2.9	14.5	0.63	2.7	13.4	0.59
9	**Access Security** Access to fetched data is restricted and hence kept secure.	5.0	0.0	25.0	4.8	23.9	0.96	3.4	17.2	0.69
	Total	43.8	--	219.0	38.2	191.1	0.90	32.1	160.6	0.70

IS-Importance Score, SD- Standard Deviation, WIS- Weighted Importance Score, RS-Raw Score, WS-Weighted Score, ICR-Intelligent Character Recognition, MDES-Manual Data Entry System, DQI-Data Quality Index

www.ijmer.com Vol.1, Issue.2, pp-270-277 ISSN: 2249-6645

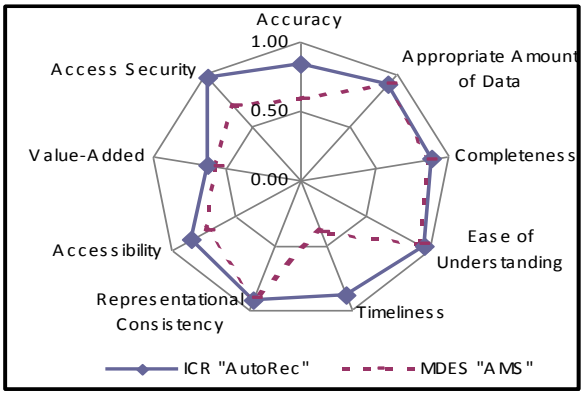

Fig 1. Input data quality dimensions or ICR and manual data entry system

5.1 EFFECT OF SCANNING PARAMETERS ON CHARACTER RECOGNITION, HUMAN INTERVENTION, COST AND TIME INVOLVED

The major difficulties of character recognition have to do with locating and correcting errors. The quality of a document's appearance is critical. A paper document marred by smudges, fingerprints, dot-matrix print, or fuzziness can be nearly as disastrous as skewed placement or a dirty scanner glass. Other errors are caused by coloured inks or papers, outsized or otherwise unrecognized fonts, etc. [2].

A good quality document with well-delineated text is the first requirement to get accuracy. Paper colour and type also affect the quality of the scanning and resulting image [1]. It was observed that different scanning parameters (threshold and contrast values) affected the recognition accuracy of ICR system. Higher character recognition accuracy rate (89.56 percent) was observed using 184 units threshold and 128 units contrast value whereas low character recognition rate (86.29 percent) was obtained using 128 units threshold and 144 units contrast value. The fig. 2 shows the effect of scanning quality on recognition accuracy level of "AutoRec" system.

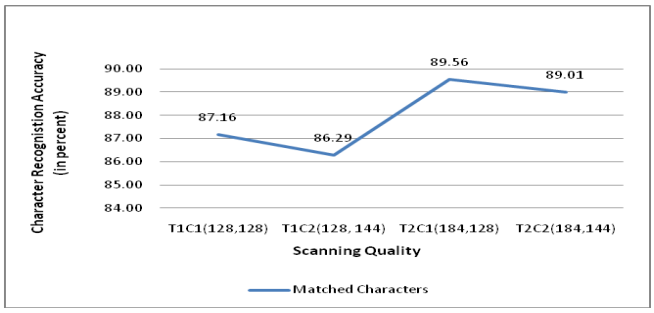

FIG.2. Effect of scanning quality on recognition accuracy level of "AutoRec" System

The Fig 3 shows the effect of scanning parameters on different errors types using "AutoRec" system. The highest errors rate (10.94 percent) for unrecognized characters is for the T1C2 images and lowest rate (7.21 percent) for T2C1 images. Similarly, highest substitution character error rate (3.55 percent) for the T1C1 images whereas lowest substitution character error rate (2.51 percent) for T2C2 images. This indicates that good combination of threshold and contrast values are required to enhance character recognition level of "AutoRec" system.

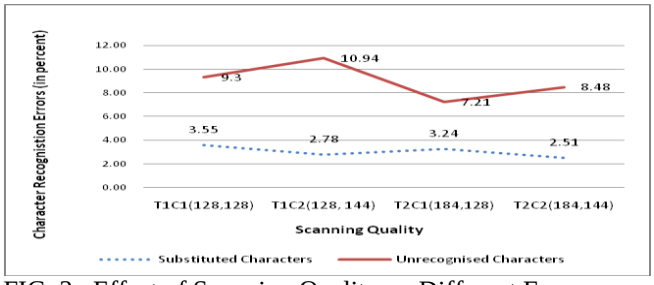

FIG. 3 Effect of Scanning Quality on Different Errors Types

The human intervention is manual efforts required at the end of the computer operator to make each individual character understandable to the computer system where "AutoRec" system is not able to recognise characters. Overall human intervention per character using "AutoRec" system was 6.0 percent whereas cost was just 0.06 units. This means for 100 characters to be entered by an operator manually, only 6 characters needs human intervention using "AutoRec" system. Similarly, if cost for manually entered 100 characters is 100 units then using "AutoRec" the cost for the same number of characters would be just 6 units. As for as promptness of "AutoRec" system is concerned, it was able to fetch 17 characters at a given time as compared to one character entered manually. It was also observed that different scanning parameters also affected the human intervention involved in "AutoRec" system. The maximum human intervention (6.86 percent) was involved for the T1C2 scanned images and minimum human intervention (5.22 percent) for T2C1 scanned images. It is concluded that the ICR based "AutoRec" system provides better performance as compared to manual data entry system in terms of time, cost and involvement of minimum human

intervention. The table 2 shows the performance of "AutoRec" system over manual data entry system using different scanning parameters.

TABLE 2
Performance of "AutoRec" System over Manual Data Entry System using Different Scanning Parameters

Images	TC	SC	UC	MC	HI (in % age)	Prm.	C
T1C1	6088	216	566	5306	782 (6.42)	15.57	0.06
T1C2	6088	169	666	5253	835 (6.86)	14.58	0.07
T2C1	6088	197	439	5452	636 (5.22)	19.14	0.05
T2C2	6088	153	516	5419	669 (5.49)	18.2	0.05
Total	24352	735	2187	21430	5.99	16.87	0.06

TC- Total Characters, **SC**-Substituted Characters, **UC**-Unrecognised Characters, **MC**-Matched Characters, **HI**- Human Intervention, **Prm**-Promptness**, C**-Cost
* Manual Data Entry involves double data entry of awards

5.2 Effect of Data Validation Checks on Character Recognition, Human Intervention, Cost and Time Involved

Data validation checks affect the recognition accuracy of ICR system. Appropriate use of various data validation checks can provide high data throughput and able to minimise human intervention. The fig. 4 shows the effect of data validation checks on recognition accuracy level of ICR system.

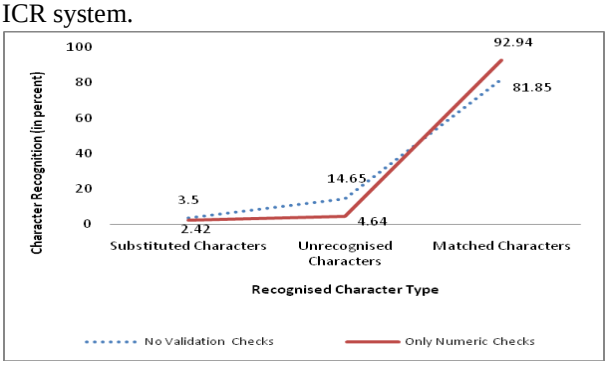

FIG. 4 Effect of data validation checks on recognition accuracy level of ICR system

The "AutoRec" system uses its own intelligence power within the domain of validation checks. It was observed that recognition accuracy level of 92.94 percent achieved using Only Numeric Checks (ONC) on award lists followed by 81.85 percent by applying No Validation Checks (NVC). Further by applying NVC, the substituted error rate and unrecognised character rate were 3.5 percent and 14.65 percent respectively whereas these were 2.42 percent and 4.64 percent by applying ONC. So, it is concluded that specific validation checks must be applied to minimise human intervention and to enhance overall accuracy of "AutoRec" system. The table 3 shows the performance of "AutoRec" system over manual data entry system using different data validation checks.

TABLE 3
Performance of "AutoRec" System over Manual Data Entry System using Different Data Validation Checks

VC	TC	SC	UC	MC	HI (in % age)	Prm.	C
NVC	12176	426	1784	9966	2210 (9.08)	11.02	0.09
ONC	12176	294	565	11317	859 (3.53)	28.35	0.04
Total	24352	720	2349	21283	3069 (6.31)	15.87	0.06

VC-Validation Checks, **TC**- Total Characters, SC-Substituted Characters, **UC**- Unrecognised Characters, **MC**-Matched Characters, **HI**- Human Intervention, **Prm**-Promptness, **C**-Cost, NVC-No Validation Checks, ONC-Only Numeric Checks
* Manual Data Entry involves double data entry of awards

Further, It was observed that the human intervention of an operator reduced by applying specific validation checks. Maximum human intervention (9.08 percent) was observed using NVC whereas minimum human intervention (3.53 percent) by applying ONC. Overall human intervention per character using "AutoRec" system was 6.31 percent whereas character recognition cost per character was just 0.06 units as compared to manually entered character. The maximum character recognition cost for ICR system was 0.09 units by applying NVC whereas minimum character recognition cost is 0.04 units on applying ONC. Further, the ICR system is 15.87 times faster to recognise characters as compared to similar number of characters punched by an operator manually in which maximum character recognition promptness (28.35 characters) using ONC whereas minimum character recognition promptness (11.02 characters) using NVC. This further indicates that "AutoRec" system has better performance as compared to manually data entry system in terms of time, cost and involvement of human intervention.

5.3 Effect of Confidence Levels on Recognition Accuracy Level, Human Intervention, Cost and Time Involved

An ICR recognition engine assigns a specific confidence value for every character to be recognized. Confidence

thresholds may be modified within the software for certain fields or characters (Phillips, 2000). The confidence level of ICR system also affects the accuracy level as well as human intervention. At confidence level 50 units, the recognition accuracy rate was 87.75 percent and error rate was 12.25 percent whereas at confidence level 75 units, the recognition accuracy rate was highest (88.14 percent) and errors rate was minimum (11.86 percent). Further, as confidence level of "AutoRec" increases above 75 units, a continuous decrease in recognition accuracy level was witnessed. The recognition accuracy rate was 88.09 percent for confidence level 90 units whereas 88.02 percent for confidence level 100 units. The fig. 5 shows the effect of confidence levels on character recognition level accuracy of "AutoRec" system.

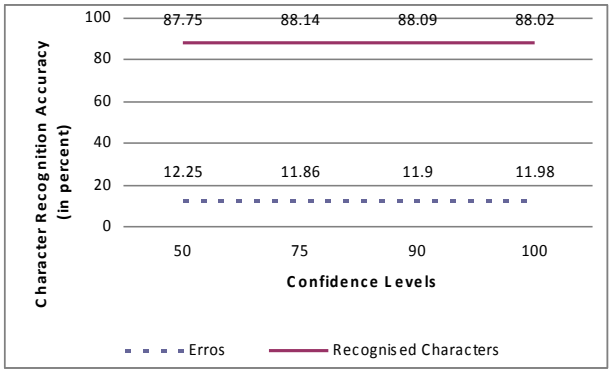

Fig. 5 Effect of confidence levels on recognition accuracy level of "AutoRec" system.

This indicates that there is a need to choose the confidence level of appropriate level to get high accuracy and minimize human intervention. It is also observed that after reaching a certain threshold limit of confidence level, there is increase in specific types of errors such as substitutional characters and these types of errors are not only very hard to detect but also very costly to correct. The fig. 6 shows the effect of confidence levels on different character recognition errors:

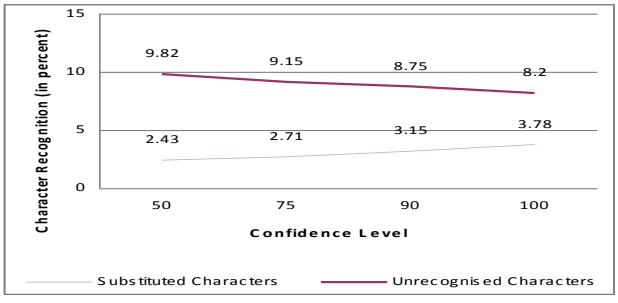

Fig. 6 Effect of confidence levels on different character recognition errors

Maximum human intervention (6.13 percent) was involved at confidence level 50 units and minimum human intervention (5.93 percent) at confidence level 75 units. The table 4 shows the performance of "AutoRec" system over manual data entry system using different confidence levels.

TABLE 4
Performance of ICR System over Manual Data Entry System using Different Confidence Levels

Confd Level	TC	SC	UC	MC	HI (in % age)	Prm.	C
50	6088	148	598	5342	746 (6.13)	16.32	0.06
75	6088	165	557	5366	722 (5.93)	16.86	0.06
90	6088	192	533	5363	725 (5.95)	16.79	0.06
100	6088	230	499	5359	729 (5.99)	16.7	0.06
Total	24352	735	2187	21430	2922 (6.00)	16.66	0.06

Confd. Level- Confidence Levels, **TC**- Total Characters, **SC**-Substituted Characters, **UC**-Unrecognised Characters, **MC**-Matched Characters, **HI**- Human Intervention, **Prm**.-Promptness, **C**-Cost
* Manual Data Entry involves double data entry of awards

The overall human intervention per character using ICR System was 6.0 percent whereas character recognition cost was just 0.06 units as compared to manually data entry system. Further, the overall character recognition promptness of ICR system was 16.66 characters as compared to one character punched by an operator manually in which maximum character recognition promptness (16.86 characters) at confidence value 75 units and minimum character recognition promptness (16.32 characters) at confidence value 50 units.

6. CONCLUSIONS

The results of the above study show that ICR technology has the potential to maintain data quality, minimise manual data entry load and increase overall productivity & efficiency of university examination system where limited human manpower, time and cost are the major constraints to process huge volume of data using papers. But to make effective utilisation of this technology, there is a need to take care of certain factors which affect the recognition accuracy level of ICR system such as quality of scanned image, use of data validation checks and confidence levels. The balanced usage of data validation checks and confidence levels do not only facilitate in minimisation of human intervention, reduction in cost and time but also increases the overall data quality. ICR system functions on individual character basis and not on entire data field so a single false recognition or substitutional error has a very

high probability to corrupt the whole record and which in turn provides poor data quality. Further, ICR system is not the substitute of human operator but an aid to minimize manual intervention for conversion of data available on papers into computer readable form. Based on the study presented in this paper, it is recommended that usage of ICR based technology "AutoRec" can be extended to other areas where huge volume of paper work is involved such as for admission, examination, settlement of result discrepancies, re-evaluation cases, etc. In addition to above, the usage of ICR technology can be used to create data centres for universities where paper is still the dominant media for exchange of information.

7. REFERENCES

[1] (2005). "The Importance of Power/Precision Data Entry to Document Imaging". [Online]. Available: http://www.vikingsoft.com/pdf/importanceofppde.pdf

[2] Arrington, Daniel V. (1992). "Departmental Document Imaging: Issues and Concerns." [Online]. Available: http://cool-palimpsest.stanford.edu/bytopic/imaging/depimgng.html.

[3] Gingrande, Arthur. "Forms Automation: From ICR to E-Forms to the Internet". *Silver Spring, MD: AIIM International*.1998.

[4] O'Gorman Lawrence and Rangachar Kasturi, "What is a Document Image and What Do We Do With It." *Document Image Analysis. IEEE Computer Society Executive Briefings*. [Online]. Available: http://www.ijcaonline.org/volume10/number5/pxc3871991.pdf

[5] Phillips, JOHN T (2000). "Does ICR Keep Paper Forms Viable?" [Online]. Available: https://www.entrepreneur.com/tradejournals/article/62194277.html

[6] Wang, Richard, Diane Strong, and Lisa Guarascio. "An empirical investigation of data quality dimensions: A data consumer's perspective." Working paper *TDQM-94-01, MIT TDQM Research Program, E53-320, 50 Memorial Drive*, Cambridge.1994.