International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 595-599

Intelligent Character Recognition- An Emerging Pattern Recognition Technology to Minimise Human Intervention in University Examination System

Vikas Sharma[1] & Anjana Sharma[2]

The university examination system is a complex system which passes through mainly three phases- pre-examination, post-examinations and post result declaration. Huge volume of different categories of data is involved in all above phases starting from fetching of personal details of the candidates to posting of awards for compilation of results to settlement of discrepancy cases. Manual data entry from handwritten documents is not only time consuming but also a very tedious, costly and prone to errors especially where huge volume of handwritten data is involved. In post-examination phase, awards lists play very vital role in compilation of results. Numerous forms based Intelligent Character Recognition (ICR) Systems have appeared as the potential solution to minimise human intervention, improve efficiency, accuracy and reduce cost while fetching handwritten data directly into machine readable form. In this paper, an attempt is made to analyse the cost, time and human intervention involved in manual data entry system viz.-a-viz. ICR system "AutoRec" implemented in Examination Wing of Himachal Pradesh University, Shimla using different scanning parameters, data validation checks and confidence levels. It has emerged that depending on the local hardware conditions and ICR software, a balance needs to be maintained to minimise human intervention, cost and time while using different scanning parameters, validation checks and confidence level values.
Keywords: ICR, Threshold Value, Confidence Level, Data Validation Checks, Human Intervention, Cost, Time.

1. Introduction

The colleges and universities throughout the country are struggling to find some optimal way to deal with their paper based documents to ensure institutional credibility and accountability. The cost of storing, filing, and finding documents has continuously been rising even after adoption of modern technologies[1]. Computer understands alphanumeric characters as American Standard Code of Information Interchange (ASCII) typed on a keyboard where each character or letter has a unique code. However, computer itself cannot distinguish characters and words from the scanned images of paper documents. Therefore, an application of Artificial Intelligence is used to match images of characters available on scanned documents and convert them into their ASCII equivalents to get readable text [3]. Document imaging describes a process whereby sheets of paper are passed through a page scanner to produce graphic images or pictures. Imaged document files (images) are processed with the aid of appropriate character recognition software to retrieve, print, alter or store as database file [1]. Numerous possibilities exist for errors in character recognition. Document scanners can misread an image that is dirty or too skewed. Characters squashed together may be interpreted as a different, larger character. Characters read without contextual analysis may be interpreted as letters, when only numbers should exist in a field [2]. There are two types of errors, rejection and substitution. In rejection the system is unable to read the character and does not recognize the character at all. With substitution error, the system miss recognizes the character and substitutes the character with high level of confidence [4]. The character recognition accuracy cannot be 100 per cent since there is variation in human handwriting depending on mood, stress and other external factors [1]. The ICR engines can achieve very high recognition rates when the documents are properly designed, printed and controlled. Nonetheless, a 1% error rate on a printed page with 3,000 characters means there is an average of 30 errors on each page. This would be unacceptable for a typist, but may be adequate for information that will only be read occasionally [5]. The ICR technology seems to be a good facility for human operators to minimise their data entry time and workload and increase overall productivity. In this paper, a comparative study is conducted to analyse the cost, time and human intervention involved in manual data entry system viz.-a-viz. ICR technology implemented for automation conversion of awards from manual handwritten awards lists to computer readable form for processing of results.

2. Research Objectives

1. To analyse the effect of image quality, confidence levels and validation checks on cost, time and human intervention involved in ICR system.
2. To compare the cost, time and human intervention involved in manual data entry system viz.-a-viz. ICR technology.

[1]International Centre for Distance Education and Open Learning, Himachal Pradesh University, Summer Hill, Shimla, INDIA
[2]Web Tech Computer Education, Sanjauli, Shimla, INDIA
Email: [1]vikas_doeacc@rediffmail.com, [2]anjana_vashisht@yahoo.com

3. RESEARCH METHODOLOGY

Five samples of ICR compatible awards lists of B. Com. 3rd Year Examinations, September 2008 of Himachal Pradesh University, Summerhill, Shimla were selected using a convenient sampling technique. The sample had 208 data fields (203- numeric data fields and 5 alphanumeric data fields) which in turn summed to recognition of 761 characters. The data types of roll number, awards and serial number field was of type numeric whereas paper code had alphanumeric. The cost, time and human intervention involved in ICR system ("AutoRec") were studied using different scanning parameters, data validation checks and confidence levels as follows:

1. Four scanned image cases (T1C1, T1C2, T2C1 and T2C2) were designed using different scanning parameters (threshold –'T' and contrast- 'C' values). The table 1 shows the scanned image cases using different threshold and contrast values:

Table 1
Scanned Image Cases Using Different Threshold and Contrast Values

Sr. No.	Case No.	Threshold Value (units)	Contrast Value (units)
1.	T1C1	128	128
2.	T2C1	184	128
3.	T1C2	128	144
4.	T2C2	184	144

2. Two types of validation checks namely: 1) AVC (All Validation Checks) –numeric, alphanumeric, letters & special characters, and 2) NVC (Numeric Validation Checks) were applied on all the above four cases (T1C1, T1C2, T2C1 and T2C2) separately.

3. Four confidence values (50, 75, 90 and 100 units) were tested separately using AVCs and NVCs in Solid Valid Section of ICR system.

4. The opinion of ICR system on character recognition level in Solid Valid Section of "AutoRec" system was classified using five point Likert Scale as mentioned in table 2.

Table 2
Classification of Intelligent Character Recognition System's Opinion

Sr. No.	ICR Response/ Opinion	Description
1.	Highly Mismatched Character	Substitution Errors/Extra Addition of characters.
2.	Mismatched Character	Incorrect recognition such as recognition of '2' as '7' or '3' as 'B', etc.
3.	Undecided Character	Unrecognised Character marked as '?'
4.	Matched Character	Recognised character but needs human intervention for its validation.
5.	Highly Matched Character	Complete match by ICR system without any human intervention.

5. The response of ICR system based on above five point Likert Scale was divided in to two segments- characters needed human intervention (Highly Mismatched, Mismatched, Undecided and Matched characters) and characters needed no human intervention (Highly Matched Character). The total number of characters needed human interventions were compared with actual manual data entry required in manual system to analyse the cost and time involved in both systems. In manual data entry system, double entry of every single award is done to bring accuracy and to avoid any kind of discrepancy.

6. Four separate observers recorded the opinions of ICR System on Solid Valid Stations of "AutoRec" system independently using network environment system. The 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.

4. RESULTS AND DISCUSSION

4.1. Effect of Scanning Parameters

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 ICR system has certain doubts on recognition of character using scanned images. These characters are classified as Highly Mismatched, Mismatched, Undecided and Matched Character. Overall human intervention per character using ICR System is 34.42 percent whereas cost is just 0.17 units as compared to one character entered manually by an operator. Further, the ICR system is 5.81 times faster to recognise characters as compared to one character punched by an operator manually. It was observed that different scanning parameters affected the human intervention involved in ICR system. The maximum human intervention (35.23 percent) was involved for the T1C1 scanned images and minimum human intervention (33.62 percent) for T1C2 scanned images. The ICR system recognised characters 5.92 times faster from scanned images (T1C2) as compared to other scanned images. The cost of ICR system was 0.17 units for (T1C2) scanned images as compared to other images. This indicates that T1C2 scanned images provides fast recognition of characters by involving minimum human

intervention and cost. It is concluded that the ICR system provides better performance as compared to manually data entry system in terms of time, cost and involvement of less human efforts. The table 3 shows the performance of ICR System over manual data entry system using different scanning parameters.

Table 3

Performance of ICR System(Human Intervention, Promptness and Cost) per Character over Manual Data Entry System using Different Scanning Parameters

Scann-ed Image Cases	TC#	HMSC#	MSC#	UC#	MC#	HM#	Human Interve-ntion Using ICR System (in percent) a+b+c+d	ICR Character Recognition Promptness Over Manual Data Entry* (per character)	ICR Usage Cost Over Manual Data Entry (per character)
		(a)	(b)	(c)	(d)	(e)			
T1C1	6088	138	566	78	1363	3943	2145 (35.23)	5.68	0.18
T1C2	6088	123	516	30	1378	4041	2047 (33.62)	5.95	0.17
T2C1	6088	169	439	28	1417	4035	2053 (33.72)	5.93	0.17
T2C2	6088	152	666	17	1302	3951	2137 (35.10)	5.70	0.18
Total	24352	582	2187	153	5460	15970	8382 (34.42)	5.81	0.17

#TC- Total Characters, HMSC-Highly Mismatched Characters, MSC-Mismatched Characters,

UC-Undecided Characters, MC-Matched Characters, HMC- Highly Matched Characters

* Manual Data Entry involves double data entry of awards by two different operators to bring accuracy in data.

4.2. Effect of Confidence Values

It was also observed that as the confidence level of ICR system increases, the required human intervention reduces. Maximum human intervention (42.90 percent) is involved at confidence level 50 units and mimimum human intervention (28.47 percent) at confidence level 100 units. The overall human intervention per character using ICR System is 34.42 percent whereas character recognition cost is just 0.17 units as compared to one character entered manually by an operator. The maximum ICR character recognition cost (0.21 units) was observed for confidence value 50 units whereas minimum ICR character recognition cost (0.14 units) for confidence value 100 units. Further, the overall character recognition promptness of ICR system is 5.81 characters as compared to one character punched by an operator manually in which maximum character recognition promptness (7.03 characters) at confidence value 100 units and minimum character recognition promptness (4.66 characters) for confidence value 50 units. The table 4 shows the performance of ICR System over manual data entry system using different confidence values.

4.3. Effect of Validation Checks

Using different data validation checks on ICR system, it was observed that on specifying the data domain types for different data fields, the human intervention of an operator had also reduced. Maximum human intervention (36.85 percent) is involved using all data validation checks whereas minimum human intervention (31.99 percent) for only numeric checks. The overall human intervention per character using ICR System is 34.42 percent whereas character recognition cost is just 0.17 units as compared to one character entered manually by an operator in which maximum character recognition cost for ICR system is 0.18 units for applying all validation checks on data fields whereas minimum character recognition cost is 0.16 units on applying only numeric checks.

Further, the ICR system is 5.81 times faster to recognise characters as compared to one character punched by an operator manually in which maximum character recognition promptness (6.25 characters) using only numeric checks whereas minimum character recognition promptness (5.43 characters) using all validation checks. This also indicates that ICR system performs better as compared to manually data entry in terms of time, cost and involvement of less human intervention when different data domains types are defined on specific data fields. The table 5 shows the performance of ICR system over manual data entry system using different data validation checks.

Table 4

Performance of ICR System(Human Intervention, Promptness and Cost) per Character over Manual Data Entry System using Different Confidence Values

ICR Confi-dence Values	TC#	HMSC#	MSC#	UC#	MC#	HM#	Human Interve-ntion Using ICR System (in percent) a+b+c+d	ICR Character Recognition Promptness Over Manual Data Entry (per character)	ICR Usage Cost Over Manual Data Entry (per character)
		(a)	(b)	(c)	(d)	(e)			
50	6088	107	598	41	1866	3476	2612 (42.90)	4.66	0.21
75	6088	126	557	39	1404	3962	2126 (34.92)	5.73	0.17
90	6088	161	533	31	1186	4177	1911 (31.39)	6.37	0.16
100	6088	188	499	42	1004	4355	1733 (28.47)	7.03	0.14
Total	24352	582	2187	153	5460	15970	8382 (34.42)	5.81	0.17

#TC- Total Characters, HMSC-Highly Mismatched Characters, MSC-Mismatched Characters,

UC-Undecided Characters, MC-Matched Characters, HMC- Highly Matched Characters

* Manual Data Entry involves double data entry of awards by two different operators to bring accuracy in data.

Table 5

Performance of ICR System(Human Intervention, Promptness and Cost) per Character over Manual Data System Using Different Data Validation Checks

Valida-tion Check Types	TC#	HMSC#	MSC#	UC#	MC#	HM#	Human Interve-ntion Using ICR System (in percent) a+b+c+d	ICR Character Recognition Promptness Over Manual Data Entry (per character)	ICR Usage Cost Over Manual Data Entry (per character)
		(a)	(b)	(c)	(d)	(e)			
AVC[1]	12176	373	1622	68	2424	7689	4487 (36.85)	5.43	0.18
ONC[2]	12176	209	565	85	3036	8281	3895 (31.99)	6.25	0.16
Total	24352	582	2187	153	5460	15970	8382 (34.42)	5.81	0.17

#TC- Total Characters, HMSC-Highly Mismatched Characters, MSC-Mismatched Characters,

UC-Undecided Characters, MC-Matched Characters, HMC- Highly Matched Characters

* Manual Data Entry involves double data entry of awards by two different operators to bring accuracy in data.

5. CONCLUSIONS AND RECOMMENDATIONS

The results show that ICR technology has the potential to minimise manual data entry load and increase overall productivity & efficiency of university like institutions where time and cost are the major constraints. It has been cleared that certain quality parameters of scanned images affected the recognition accuracy level of ICR system and required human efforts to make data compatible to computer system for further its further processing, etc. The balanced use of data validation checks and confidence levels do not only facilitate in minimisation of human intervention but also helpful in reduction of time, cost and overall increase in data fetching accuracy. The data validation checks should be used appropriately in ICR system wherever needed for speedy recognition and to achieve high data quality. Using confidence values, the ICR system adds assurance level and skips some characters based on doubts during recognition process. But use of high confidence level increases the false character recognition or substitutional errors. These types of error are very hard to detect and need human intervention

for correction. So in nutshell it is concluded that there is a close dependence of scanner properties, data validation checks and confidence levels in ICR system to fetch accurate data by involving minimum human intervention, cost and time.

Based on the study presented in this paper, it is recommended that use of any ICR system is reliable provided properly tested scanning parameters, validation checks and confidence values are coherent to each other. The use of ICR technology can be extended to fetch students' personal information such as name, father's/mother's name, appearing paper details, examination centres, etc. including their photographs and signatures directly from their ICR compatible admission or examination forms in addition to rapid settlement of result discrepancies and re-evaluation cases for university examination systems. The balance use of scanner parameters, data validation checks and confidence levels can definitely increase the overall accuracy of data and productivity of examination systems by bringing down the overall cost of data entry and additional requirement of human efforts for collation of results manually.

REFERENCES

[1] Arrington, Daniel V, "Departmental Document Imaging: Issues and Concerns", 2008, February 17, 2009 <http://cool-palimpsest.stanford.edu/bytopic/ imaging/ depimgng.html >.

[2] IMJ, "Does ICR Keep Paper Forms Viable?". April 2000. Information Management Journal 17 Aug. 2009. <https://www.entrepreneur.com/ tradejournals/article/62194277.html>

[3] Parascript, "ICR Software". 2009.< http://www.parascript.com/company2/icr_software.cfm>

[4] Rosen, Richard, Vinod Kapani, Patrick Clancy, "Data Capture using FAX and Intelligent Character and Optical Character Recognition (ICR/OCR) in the Current Employment Statistics Survey (CES)", 2003 Federal Committee on Statistical Methodology. 2009 <http://www.fcsm.gov/03papers/Rosen.pdf>.

[5] Viking "The Importance of Power/Precision Data Entry to Document Imaging, A White Paper". 2005. Viking Software Solutions. 2009. <http://www.vikingsoft.com/pdf/importanceofppde.pdf>