ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Handwritten Document Management System: Key Challenges And Probable Solutions For Indian Railway And Healthcare Industries

Sandip Rakshit
Army Institute of Management
Kolkata, India
rakshitsandip@ieee.org

Kalyan S Sengupta
Indian Institute of social welfare and Business Management
Kolkata, India
kalyansen@iiswbm.edu

Subhadip Basu*
CSE Department, Jadavpur University
Kolkata, India
subhadip@ieee.org

Abstract: In this paper we propose a Handwritten Document Management System (HDMS) which manages handwritten paper documents and their applications in Indian Railway and Healthcare Industries. The proposed HDMS manages paper documents, including those with handwriting, and integrates them with electronic documents. By digitizing and managing handwriting on paper, we provide document management and retrieval capabilities that utilize the OCR technology. In reality, handwritten data usually touch or cross the pre-printed form frames and texts, creating complex problems for the recognition routines. In this paper, we address these issues and attempted to solve the problem for two real-life problem domains using our custom built form processing software and Tesseract open source character recognition engine.

Keywords: Handwritten Document Management System, Tesseract OCR engine, Technology Acceptance Model

I. INTRODUCTION

A Document Management System (DMS) is used to track and/or store digitized paper documents in a repository. It may be also termed as Document Information Systems (DIS), Integrated Document Management (IDM) system, Electronic Document Management (EDM) system, and Enterprise Content Management (ECM) system etc., based on the actual application domain. DMS is a process taken with the documents within an organization, with respect to the creation, distribution and deletion of such documents. It has been estimated that 90% of an organization’s information is in documents rather than in structured databases [1].

Powerful indexing/retrieval techniques need to be integrated in DMS. The need for searching archives of scanned handwritten documents is required in applications such as historical manuscripts, forensic/criminal records, personal documents etc. Need based applications for various business segments like healthcare, Railways, financial/retail sectors etc. can also be developed with support from DMS. For all such applications there is always a need for indexing and retrieval based on textual contents and user-indexed keywords

Document management system components includes capture, process, manage, preserve, deliver, indexing, storage, retrieval, recognition, search, etc. The term capture includes functionalities for generating, capturing, preparing and processing of electronic information. In process step various recognition technologies are used to recognize such data. Then we index all those data in methodical way. In manage step we put the data in database for administration and retrieval step fetches necessary information from the DMS.

The main benefits and features of the proposed Handwritten Document Management System (HDMS) include a centralized repository for storing, version management, audit logs to track creation, Modification and deletion, restricted access etc. Paper documents can be scanned and stored in the repository. Extensive indexing options enable easy and faster retrieval and sharing of documents [2-5].

II. REVIEW OF PREVIOUS WORK

In an early work on computerized prescribing, Schiff et al. [6] explains that the electronic prescription management system could have a positive impact on drug selection; the patient role in pharmacotherapy risk-benefit decision making; screening for interactions (drug-drug, drug laboratory, and drug-disease); linkages between laboratory and pharmacy; dosing calculations and scheduling; coordination between team members, particularly concerning patient education; monitoring and documenting adverse effects; and post marketing surveillance of therapy outcomes.

Jain et al. [7] proposed a string matching based method for word spotting in on-line document. The on-line data are collected through the help of various electronic devices like PDAs and notebook PCs. Word-spotting is based on a direct comparison of a handwritten keyword to words in the document and it is also used for indexing and retrieval purpose.

Sasche et al. [8] described various techniques for searching in handwritten documents using approximate string searching. For textual recognition they used stroke direction features, biometric features, fusion of single-feature etc. These techniques were used for finding parts within sequences of amino acids or genes.

In one of our earlier works [9] the concept of HDMS has been applied in the context of Indian railway reservation/cancellation requisition system with encouraging results. In reality, handwritten data usually touch or cross the preprinted form frames and texts, creating complex problems for the recognition routines. In this paper, we address these issues and attempted to solve the problem for Indian Railway Reservation system using our custom built form processing software and Tesseract open source character recognition engine.

Basu et al. [10], designed a novel query retrieval scheme for the information just in time (iJIT) system to retrieve handwritten annotations from digital documents based on typed/handwritten query. The key components of this query retrieval system (QRS) are the character recognition engine and the query retrieval engine. The character recognition engine receives real-time digital pen generated data, and produces segmented-recognition result. The query retrieval engine, resolves the index/query requests from the users for possible information update/retrieval. In case of a handwritten query, the query retrieval engine interacts with the recognition engine to create/update the inverted index table with recognized word labels with annotation indices. In the case of typed text query, the inverted index table is searched directly to retrieve the best matches of annotation indices using a q-gram based approximate string matching technique.

III. DESIGN OF THE HDMS

The objective of the current work is to develop a scheme to integrate different components of a handwritten document management system and automate the recognition and retrieval steps involved therein. A structured bloack-diagram of the overall system is shown in Fig.1. As proposed in our system, the input to the institutional data-server can be made in one of four ways, 1) through manual entry into the system, 2) via web-based interface, where the users/operators enter data into the system in a client/server environment, 3) through a custom-designed graphical user interface, where the scanned/digitized hand-filed data sheets/documents are processed using an intelligent OCR system and 4) through an instant digitization device, coupled with a smart data-processing algorithm for instant recognition/indexing into the data server. In the current work we have interest in the last two modes of data acquisition methodologies. We propose a general solution for HDMS by redesigning the input data-sheets, integrating the OCR technology for automatic recognition of isolated handwritten text images and enabling automatic digitization of handwritten annotations. Technicalities involved in recognition of isolated handwritten characters in a custom HDMS has already been reported in [11-14]. An automatic indexing and retrieval technique for handwritten annotations in a just-in-time system was also reported in one of our earlier works [10]. The current work proposes the overall integration scheme, with case studies in two mutually disjoint application domains.

One of the most vital components in our HDMS is the design of the pattern classifier for recognition of handwritten text patterns. We have used the Tesseract open source OCR engine [15] for pattern classification and recognition. Tesseract is an open source (under Apache License 2.0) offline optical character recognition engine, originally developed at Hewlett Packard from 1984 to 1994. Tesseract is now partially funded by Google [16] and released under the Apache license, version 2.0. The latest version, Tesseract 2.03 is released in April, 2008. In the current work, we have used Tesseract version 2.01, released in August 2007.

Like any standard OCR engine, Tesseract is developed on top of the key functional modules like, line and word finder, word recognizer, static character classifier, linguistic analyzer and an adaptive classifier. However, it does not support document layout analysis, output formatting and graphical user interface. Currently, Tesseract can recognize printed text written in English, Spanish, French, Italian, Dutch, German and various other languages.

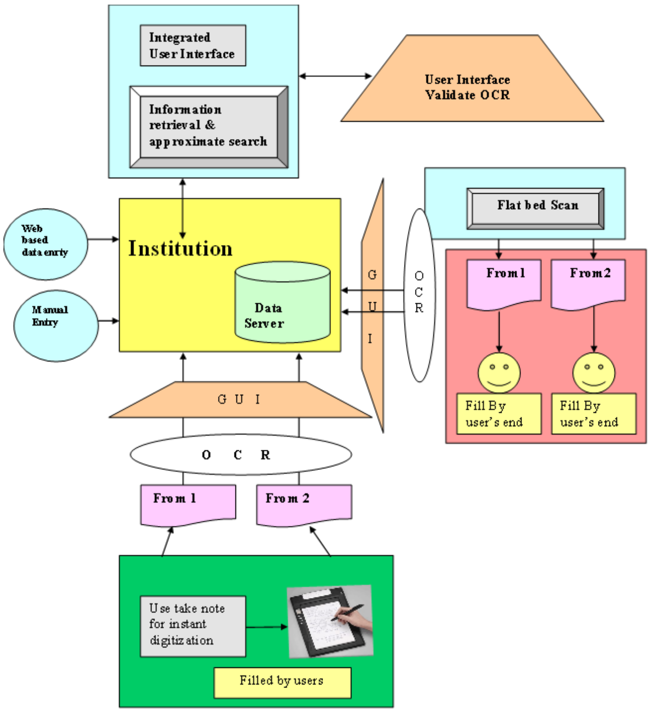


Figure. 1. A schematic diagram of the overall handwritten document management system is shown.

To train Tesseract in English language 8 data files are required in tessdata sub directory. The 8 files used for English are to be generated as follows:

tessdata/eng.freq-dawg
tessdata/eng.word-dawg
tessdata/eng.user-words
tessdata/eng.inttemp
tessdata/eng.normproto
tessdata/eng.pffmtable
tessdata/eng.unicharset
tessdata/eng.DangAmbigs

With the help of the above classifier we build the proposed HDMS, which is validated over two separate application domains, one for the Indian Railways Reservation System (IRRS) and the other for the Health Care Automation System (HCAS).

A. *Case Study1: IRSS:*

Indian Railway is the largest rail network in Asia and the world's second largest under one management. It covers 64000 of route kilometre along length and width of country [17]. It runs 12000 trains every day, it also carry 14 million passengers and 2 million tonnes of goods every day. Our main concentration of the current research is to re-design the current railway reservation or cancellation requisition form and recognize the contents of the specified forms in various aspects. Currently there are two different options to book or cancel tickets in Indian railway, i.e., through on line or web based interface and using manual ticket booking system. Despite increasing popularity of the web based (online) ticket reservation or cancellation system, vast majority of the Indian population still uses paper based reservation requisition forms for this purpose. Fig 1(a) shows an example of such form, now in use, both in English and in different Indian languages. Presently Indian Railway uses unstructured forms (as shown in Fig 1(a)) that are entered manually into the system by the on-duty counter clerk, at the Indian Railway Reservation counters. It is time consuming both for the clerk and for the customers waiting in long queues in such counters. To address this problem we have designed a structured reservation/ cancellation requisition form for IRRS, as shown in Fig 1(b) that can be processed automatically by our designed system.

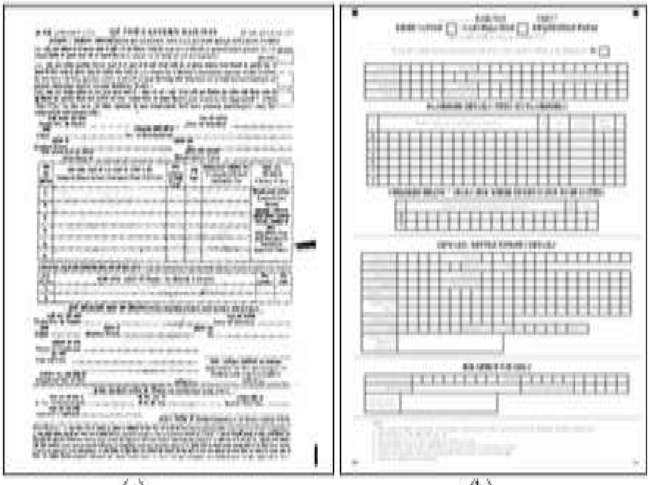
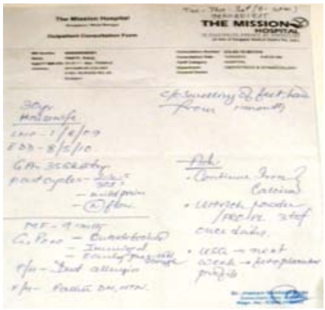


Figure 2. Scanned images of the IRRS forms are shown. (a) Original Railway reservation/ cancellation form, currently in use in India. (b) The structured form designed for our proposed IRRS system.

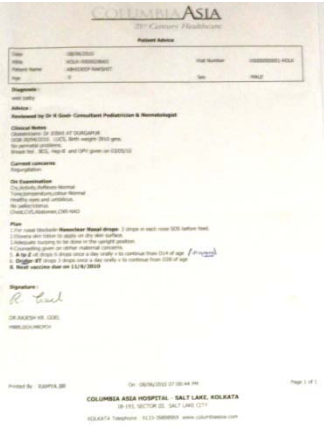
B. Case Study2: HCAS (Proposed):

Physicians and other medical professionals are slowly embracing the advantages of using information technology in their practices. Some sources estimate that only 5 percent of physicians use electronic medical-record systems [22]; the remaining 95 percent of patient medical records are paper-based. Initiatives such as HIPAA (Health Insurance Portability and Accountability Act of 1996), the availability of easy-to-use inexpensive technology, and a more technologically savvy and informed patient base are likely to bring rapid changes to the medical community.



THE MISSION HOSPITAL

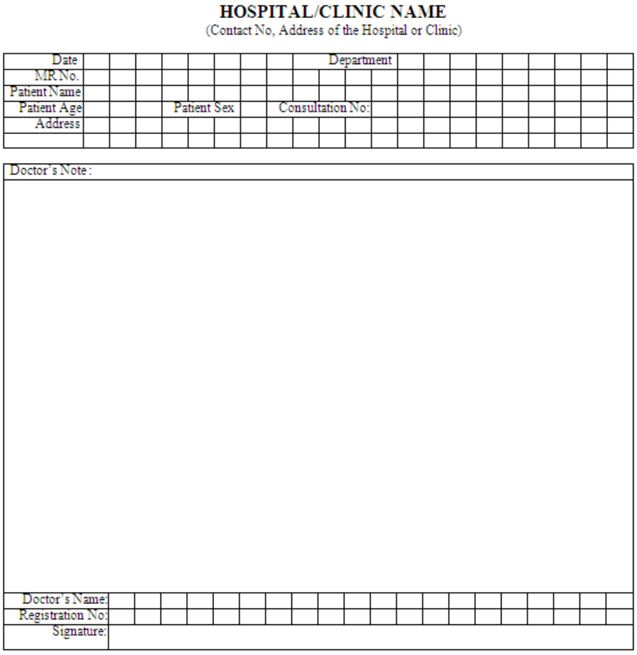
(a)



COLUMBIA ASIA

COLUMBIA ASIA HOSPITAL - SALT LAKE, KOLKATA

(b)



HOSPITAL/CLINIC NAME

(Contact No, Address of the Hospital or Clinic)

Date		Department	
MR No.			
Patient Name			
Patient Age	Patient Sex	Consultation No:	
Address			

Doctor's Note :

Doctor's Name:	
Registration No:	
Signature:	

(c)

Figure. 3. (a-b) Digitized images of unstructured medical prescriptions, (c) a structured prescription format designed for the proposed HDMS.

Document Management in a hospital as part of a health care system may be proposed. In such an application, a large variety of Handwritten documents(Prescriptions, Patient case history, Diet- charts, etc) are used along with typewritten or digital documents like pathological reports, patient bills and receipts, patient's feedback etc. In the coming years, the digitization of information and the Internet will be extremely powerful in reducing healthcare costs while assisting providers in the delivery of care. One example of healthcare inefficiency that can be managed through information digitization is the process of prescription writing. Due to the handwritten and verbal communication surrounding prescription writing, as well as the multiple tiers of authorizations, the prescription drug process causes extensive financial waste as well as medical errors, lost time, and even fatal accidents. Electronic prescription management systems are being designed to address these inefficiencies. By utilizing new electronic prescription systems, physicians not only prescribe more accurately, but also improve formulary compliance thereby reducing pharmacy utilization. These systems expand patient care by presenting proactive alternatives at the point of prescription while reducing costs and providing additional benefits for consumers and healthcare providers.

In the present work we have designed a structured patient prescription document for automatic extraction of handwritten text from the digitized images. Fig. 3(a-b) shows sample digitized images of unstructured medical prescriptions, and Fig. 3(c) shows a structured prescription format designed for the proposed HDMS. Please note that the large text area (Doctor's note) will be digitized as a single image and will not be recognized by the Tesseract pattern classifier. We may employ handwritten text annotation technique in future, described in [10], for automatic indexing and retrieval of the free-flow text.[18-22]

IV. FEASIBILITY ANALYSIS IN MANAGEMENT PERSPECTIVES

A number of feasibility models for various commercial applications have been developed by past researchers. Fig 4(a) describes the technology acceptance model [23] where the behavioral intensions of the users are measured for a new technology, considering its perceived usefulness. Naturally when the intensions of the users are very high there is a high probability for utilization of the system and the system becomes commercially viable.

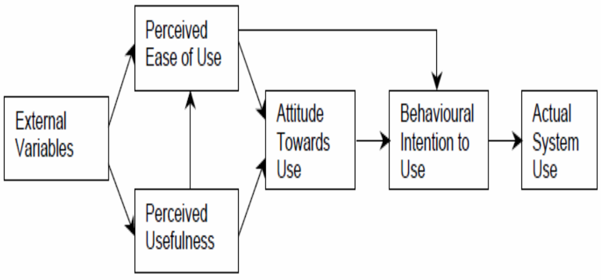
In BPA (Business Process adaptation) technology adaptations model [24] Fig 4(b), the system is evaluated from the perspective of the organization, who tries to adapt the new system. Here the process is evaluated in terms of:

a. Process –design fit
b. Technology process fit
c. Technology organization fit

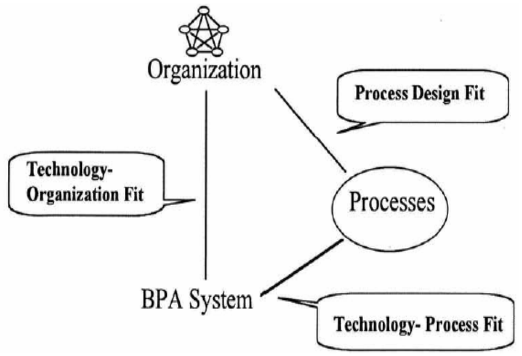
It is very much desired that the proposed business process (in our case proposed HDMS) should fit seamlessly in the organization along with its existing processes.

Capture and store all these documents in form of electronic media in such a way that these could be retrieved quickly as and when required. The proposed framework will be validated by collecting sample data from a selected real life application of HDMS. It is very important for the organizations to

The study proposes to implement such a system and also assess the attitude of the application users in order to derive the practical feasibility of such a system in a given society Handwritten document management, i.e., digitization, transmission, recognition, indexing and retrieval, is a challenging task for many commercial applications. Therefore solving this problem will be a much beneficial for the society and have a huge techno-commercial impact on most business segments



(a).



(b)

Figure 4. (a) The Technology Acceptance Model (TAM) and (b). the Triangular Model for BPA Technology Adaptation is shown.

V. DISCUSSION AND CONCLUSION

The current work is an application of HDMS in IRRS and HCAS. We also use here Tesseract OCR Engine as a pattern classifier. Instead of Tesseract we also can use another pattern classifier for better recognition and increase the accuracy of the system. On the other hand we also discuss the feasibility analysis of the said two cases. For that purposes we have taken help the two behavioural models, 1) Technology Acceptance Model (TAM) and technology adaptation model. In this paper we describe the integrated architecture of the system. The data model links information from paper, handwriting, and electronic documents together. It makes it possible to interweave searches for electronic documents and handwriting on paper documents. An effective document processing system must be able to recognize structured and semi structured forms that is written by different persons' handwriting. In this work we have developed a method and system that can process structured doctors Prescription document layout and recognize and retrieve its contents. Our approach has been applied here in the context of Indian railway reservation/cancellation requisition system and Doctors prescription Management System with acceptable performance.

VI. ACKNOWLEDGMENT

The first author is thankful to Army Institute of Management, Kolkata, India, for providing necessary supports during the research work. The second author also acknowledges The Indian Institute of Social Welfare & Business Management (IISWBM) for the support they have extended to develop and prepare the paper. The last author is thankful to the Center for Microprocessor Application for Training Education and Research (CMATER), SRUVM and PURSE projects of Computer Science and Engineering Department, Jadavpur University, for providing infrastructure facilities during of the progress of the work.

VII. REFERENCES

[1] R H. *Sprague* Jr, "Electronic Document Management: Challenges and Opportunities for Information Systems Managers". MIS Quarterly 19(1): (1995).

[2] Huaigu Cao, Anurag Bhardwaj, Venu Govindaraju "A probabilistic method for keyword retrieval in handwritten

document images" Pattern Recognition 42 (2009) 3374 – 3382.

[3] Jos´e A. Rodr´ıguez-Serrano, Florent Perronnin. "Handwritten word-image retrieval with synthesized typed queries" 2009 10th International Conference on Document Analysis and Recognition, pp 351-355.

[4] R. Milewski, V. Govindaraju and A. Bhardwaj.Automatic recognition of handwritten medical forms for search engines, International Journal of DocumentAnalysis and Recognition, 11(4):203-218, 2009.

[5] Hiroshi Sako et al., "Form Reading based on Form-type Identification and Form-data Recognition", Int. Conf. on Doc. Ana. and Recognition, Aug. 2003, Vol. 2, pp. 926-930.

[6] Schiff, Gordon D.,and Rucker, T. Donald. "Computerized Prescribing: Building the ElectronicInfrastructure for Better Medication Usage." Journal of the American Medical Association,1998; 279, 1024–1029.

[7] A. K. Jain and AM Namboodiri, "Indexing and retrieval of on-line handwritten documents," in Proceedings of IEEE ICDAR, pp. 655–659, 2003.

[8] Sascha Schimke, Claus Vielhauer: Pen-Based Retrieval in Handwritten Documents, 6th Industrial Conference on Data Mining, ICDM 2006,pp 253-257.

[9] S. Rakshit, S.Das, K. S. Sengupta, S. Basu, "Automatic Processing of Structured Handwritten Documents:An Application for Indian Railway reservation System" Published International Journal of Computer Applications 6(11):26–30, September 2010.

[10] S. Basu, K. Konishi, N. Furukawa, H, Ikeda, "A novel scheme for retrieval of handwritten textual annotations for information Just In Time (iJIT)", proceedings (CD) of IEEE Region 10 Conference (TENCON) -2008.

[11] Yaakov Navon, Ella Barkan, Boaz Ophir, "A Generic Form Processing Approach for Large Variant Templates," icdar, pp.311-315, 2009 10th International Conference on Document Analysis and Recognition, 2009.

[12] B. Yu and A. K. Jain, "A Generic System for Form Dropout", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 18, No. 11, Nov. 1996, pp. 1127-1134.

[13] Y. Belaid, et al., "Item Searching in Forms: Application to French Tax Form", Int. Conf. on Document Analysis and Recognition, Aug. 1995, pp. 744-747.

[14] H. Fujisawa, Y. Nakano, and K. Kurino, "Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis", Proc. of the IEEE, Vol. 80, No. 7, 1992, pp. 1079-1092.

[15] R. Smith. "An overview of the Tesseract OCR engine". In ICDAR'2007, International Conference on Document Analysis and Recognition, Curitiba, Brazil, Sept. 2007.

[16] http://code.google.com/p/tesseract-ocr/

[17] http://www.indianrail.gov.in/

[18] S.Rakshit, A. Kundu, M. Maity,S. Mandal, S. Sarkar, S. Basu, "Recognition of handwritten Roman Numerals using Tesseract open source OCR engine" Second International Conference on Advances in Computer Vision and Information Technology (ACVIT 2009) pp. 572-577.

[19] S. Rakshit, S. Basu, Hisashi Ikeda " Recognition of Handwritten Textual Annotations Using Tesseract Open Source OCR Engine FOR information Just in Time (iJIT) In Proc.of International Conference on Information Technology and business Intelligence (ITBI-09).

[20] S. Rakshit, S. Basu, "Development of a Multiuser Handwritten Recognition System Using Tesseract Open source OCR" in proc. of C3IT-2009, pp.240-247 .Proceedings published by Macmillan advanced Research Series, ISBN NO: 023-063-759-0.

[21] S. Rakshit, S. Basu, "Recognition of Handwritten Roman Script Using Tesseract Open source OCR Engine," in proc. of National Conference on NAQC-2008, pp. 141-145, Kolkata.

[22] S. Rakshit, D. Ghosal, T. Das, S. Dutta, S. Basu, "Development Of A Multi-User Recognition Engine For Handwritten Bangla Basic Characters And Digits" In Proc.(CD)of International Conference on Information Technology and business Intelligence(ITBI-09).

[23] Edward A. Stohr, J. Leon Zhao, "A Technology Adaptation Model for Business Process Automation," hicss, vol. 4, pp.405, 30th Hawaii International Conference on System Sciences (HICSS) Volume 4: Information Systems Track - Internet and the Digital Economy, 1997.

[24] F. Wahid, "Using the Technology Adoption Model to Analyze Internet Adoption and Use among Men and Women in Indonesiaa", The Electronic Journal on Information Systems in Developing Countries, Vol. 32, pp. 1-8, 2007.