Sophisticated Data Entry Application using Matchmaking Algorithm through Scanned Images

S. Aarthi and N. Vijay

Abstract--- *Due to the cost effectiveness and job overload in the field of data entry, the images are scanned by using front and back scanners from the paper. The images are processed using morphological image processing to reduce the noise and then intelligent character recognition method is needed to identify the handwriting and text which are written in the paper. Then the fields are identified as ontologies and the corresponding fields are matched using the ontology text matchmaking algorithm. Higher throughput is achieved by means of scanning and processing the text images. The job overload is reduced with providing machine learning algorithm.*

Keywords--- *Morphological Image Processing, ICR, Ontology Text Matchmaking Algorithm*

I. INTRODUCTION

DATA entry mechanism is used to type the texts which are given in the text papers or in CD to the given software or web page. It is a labour intensive task results in highly cost effective and time consumption. To avoid the error prone, the scanned images are taken directly into the account. Here, the data entry job is carried out by means of scanning the images of the text to be typed. When we are considering the voice recognition, the task contains speech recognition, pitch, and noise ratio calculation etc. It also involves the man power to produce the result. Then the templates are used to specify the elements to be edited. It results in the software installation cost and periodic maintenance. The process involved in this scheme is based on the following steps

Phase 1: Text recognition

- Scanning the images, the text is recognised by means of reducing the noise present in the images.
- Partition the images into sub modules to recognise each text clearly by implementing text recognition scheme.

Phase 2: Database CreationRepresenting each text in the form of tuples with the corresponding object id.

- Querying the datbase by means of query language.

Phase 3: Match making

- Matching the attributes in the corresponding fields like name, address etc. to recognise the text.
- It represents the calculation and testing phase automatically to eliminate the errors by taking inputs from the database.

Typically section 2 denotes the background; section 3 denotes the related work, section 4 denotes the proposed system.

II. BACKGROUND

In this section, the methods which are used for converting the handwriting and text pages into the text in the file or documents. For.eg. The exam particulars for students that are reported in the form of paper and then typed in the web and excel documents after some time. It takes much more time to provide the list on the web and it is cost effective. To reduce such problems, the papers that are scanned and maintained in the repository are processed using the ICR and OCR technologies.

A. Intelligent Character Recognition

ICR is used to convert the handwriting texts and cursive handwriting texts into various fonts that are needed. In this methodology the self learning system through neural networks is used to scan the images and updates the text in the database. The self learning system indicates the knowledge base with the training set. Training set is the collection of data used for the test collection. In the handwriting recognition, various parameters like Charater extraction, Character recognition and feature extraction are noticed.

B. Ontologies

Ontologies are used to conceptualise the data. The fields in the data entry may include the various texts with the same meaning. It is used relate the words with the same meaning and their relevance.No more than 3 levels of headings should be used.

III. RELATED WORK

Webbase by DMAC is one of the data entry software used to access remotely using the Microsoft remote desktop. In this software the data entry job is indicated by the SAP environment. It induces the results in the form of user's convience, speed and effectiveness. It uses the Microsft pre-XP Remote Desktop Protocol to provide the RDP connection.

Voice Recognition databases are used to recognize the voice by means of indicating the field. Although it is effective, it is time consuming. Then the voice recognition process is implemented by tuning and reshaping the frequency using repeater and regenerator. Microphone array database, Census database are the types of audio databases used in the real time application.

S. Aarthi, Master of Computer Science and Engineering, M.A.M College of Engineering, Tiruchirapalli, India. E-mail: aarthisilviya01@gmail.com

N. Vijay, Master of Computer Science and Engineering, M.A.M College of Engineering, Tiruchirapalli, India. E-mail: vijaycse53@gmail.com

There are many home based or part time jobs availabe for the data entry scheme by distributing the work to many online traders.

IV. PROPOSED SYSTEM

In this proposed mechanism, the scanned images are processed and regenerated in the preprocessor. The preprocessor changes the scanned text in the form of recognition. Small images are not recognised properly in the ICR mechanism. So the images must be provided with proper x and y axis and aspect pixel ratio. The text recognition procedure is manipulated using the ICR mechanism and stored the data in object oriented database. Each object contains unique ID which helps in the search mechanism. This ID is referred as Object ID. Each ID contain the other objects in the tuple manner.Then the process intimates the matchmaking algorithm to match the field in that id in corresponding tuples and attributes Finally repeat the process to enter the detai left of the MS Word Formatting toolbar.

A. *System Architecture*

The below architecture describes the system organisation and its mechanism.

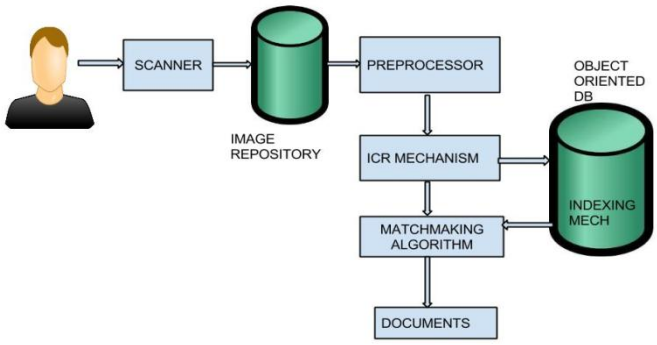

Figure. 1: System Architecture

This diagram represents the workflow from the Scanner to the changed documents. The scanned images are processed in the preprocessor to provide the clear image by reducing the noise over that image by means of morphological image processing.Then the fields in the images are partitioned and stored in the object database.

B. *Components*

The components involved in this system are scanner, Preprocessor, ICR mechanism, Matchmaking algorithm, Object Oriented Database and Image repository. They are prescribed below.

a. *Scanner*

Scanner is used to scan the documents and to convert them into the image format such as JPEG, JPG, and PNG.

b. *Preprocessor*

Preprocessor is used to eliminate the noise presnt in the images and to describe the fields clearly with the fixed partition size.

c. *Image Repository*

Image repository is used to store the images after scanned. Storing the image files becomes more complex due to the size of the files. The image size is larger when compared with the text files.

d. *ICR mechanism*

The mechanism used to convert the handwriting in the various font styles. It is the neural network helps in self learning system and knowledge bases.

e. *Object Oriented Database*

It helps in storing the text that is converted in the form of objects. Objects contains the properties such as reusability, encapsulation etc.

C. *Work Flow Analysis*

The workflow is discussed in this section. It includes the feature extraction, Character recognition, Database creation and connectivity to the database.

The scanned image for the input are described below

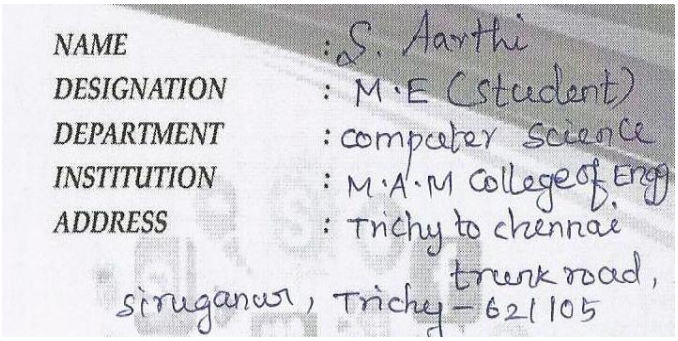

Figure. 2: Scanned image

The preprocessed image is used to enhance the image to make it clear for the the partition.

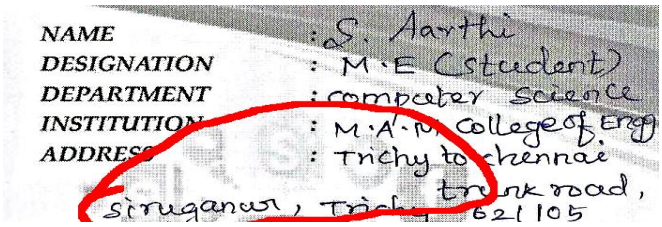

Figure. 3: Image with Noise

Due to the noise, the text is again lightened and smoothen.The background is lighten and the noise is reduced and the text is embellished

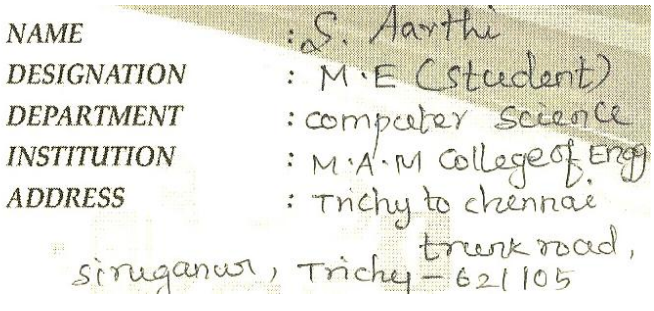

Figure. 4: Processed Image

Now the partition begins and after the ICR mechanism the result will be

S.AARTHI, M.E (STUDENT), COMPUTER SCIENCE AND ENGINEERING, M.A.M. COLLEGE OF ENGG, TRICHY TO CHENNAI TRUNK ROAD, SIRUGANUR, TRICHY-621105

Then these data stored in object oriented database in the form of tuple and a single entry with the specific object ID. It distinguigh each tuple.

The algorithm is executed to match the field using the connectivity like JDBC, ODBC etc.

D. Algorithm

a. Morphological Image Processing Algorithm

This algorithm is used in the preprocessor to reduce the noise in the scanned images. It comprises of the phases like *erosion, dilation, opening and closing phases.*

b. Erosion

It converts the irregularities present in the image. It helps in strip away the pixels to indicate whether it is optimal or not.

c. Dilation

It helps in adding the pixel layer in the image to dilute the background.

d. Opening

It is the technique in which the erosion is completed before the dilation mechanism.

e. Closing

It is the technique in which the dilation is processed before the erosion technique.

By these ways, the images are scanned and processed to reduce noise and enhance the quality of the image.

The scanned page is partitioned to recognise each text with the higher accuracy.

f. Intelligent Character Recognition

- Discover the starting point.
- Trace the contour from the starting point.
- Find the features of the contour “up”, “down”, “left”, “right”, “diagonal” and “arc”.
- Search the database whether the contour is identified or not.
- Multiple sections must induced by ICR to provide the data and to store them in specific database.

g. Ontology Text Matchmaking Algorithm

- Matchmaking is processed by means of representing each field as ontology.
- The prediction is based on finding the exact meaning of the attributes to mark in the field.
- The object that is stored in the database is queried by means simple object querying language.
- Then placing the corresponding field with the given attributes is considered.
- Connecting the database with JDBC or ODBC.
- Designation, qualification, education details are considered as the ontology with similar meaning, so it matches the attributes and insert those attributes in the web page.
- If this process includes the age calculation, total, average and verification, it manipulates automatically with testing phase.

V. Conclusion and Future Enhancement

This data entry scheme provides less time consumption and cost effective. The person who works for the data entry needs more than Rs.7 for a page. But in this system, the cost is upto 5 per page and the time consumed to provide the documents is lesser than the other data entry jobs with templates.The translation and transliteration of the text will also be added in the future enhancement to provide the support for the given system in the effective manner. By means of recognising the images with the different languages, then reproduce it with various language.

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.