DIGITIZATION AND IMAGE PROCESSING OF HISTORICAL OFFICE DOCUMENTS: AN ADMISSION SYSTEM FOR CBSE AFFILIATED SCHOOLS IN INDIA

[1]Gaurav Gupta
Research Scholar,
Department of Computer science and Engg., Dr. A.P J Abdul Kalam University, INDORE, M.P. INDIA,

ABSTRACT: ***Most of the organizations rely largely on data collection for their operations. This data is usually captured in the paper forms. Central Board of Secondary Education affiliated schools in India, uses paper admission forms to collect student data. Getting data of form and subsequently digitizing them by data entry is a traditional way to maintain a record keeping system. In this system digitizing data directly send to user who manage school database and will help them gain invaluable insights about their schools. As an alternative to paper forms online forms are gaining momentum. But online forms require wireless network access and established IT infrastructure which may not be easily available everywhere. In such cases, paper forms provide ease of use and can be distributed conveniently. Paper forms have more truth value attached to them as compared to online forms as it is easy to identify and authenticate the person filling the paper form. The problem with data collected using paper forms is aggregation and analysis. In order to convert data into machine readable form, a data entry operator is needed to feed data from paper forms into the system, which is dull and error-prone. In this paper we present a system named DMR (Digitize Manage Record). It is a desktop application developed using java technology and IDE Net Beans 8.2 with my SQL server database. It processes the image of admission form either capture by camera of any android mobile phone or from other image capture source. It accepts an image of a filled form and extracts filled information according to the template previously generated. The captured image of the empty form which is used by the DMR application to generate a template of the form.***

Keywords: DMR system, Character Recognition, Digitization, Tesseract OCR, FAST.

---

1. INTRODUCTION

The purpose of research is to present a solution that easily converts data from paper forms into a machine readable format without relying on a specific form structure. Multiple solutions are available in markets that capture data from the image of a form. But the main problem is that we need the form structure to be in a predefined format. For example, Optical Mark Recognition (OMR) technology can work only with bubble-based form specifically designed for use with the system and is unable to digitize data in existing forms. Also it needs a specialized scanning device, thus all forms have to be manually collected at one location adding physical effort. This calls for a robust solution that digitizes data from, forms of any structure or format and eliminates the need of a specialized scanning device while simultaneously providing portability.

2. RELATED WORK

A. Leadtools:

Leadtools have developed various Software Development Kit (SDK) tools that help in recognizing and processing the form. Recognition means that the document is a known form and in processing we extract information from specified areas. It uses two types of forms for form recognition: master and filled. Master forms are the blank templates that define where data is to be extracted from. After the customers receive, complete and submit their filled forms, these forms are then matched against the master forms (i.e. form recognition or classification) and then the data is extracted (i.e. form processing). It recognizes form based on any of the three characteristics: default (lines, shapes), title of the form (OCR) or barcode. The drawback of this system is that it is OCR dependent and thus portability hits.

B. CAM

CAM is a UI toolkit that allows a smartphone to interact with paper forms. Visual codes are contained in the system which act as references to help the user in communication with a remote data server and entry of data. Although CAM is a powerful tool that works well with different data types, we still need the structure of the form to be in some defined format.

DMR does not impose such restriction on the structure of the form. It provides a truly robust solution which can be easily configured to work with any type of form.

C. *QueXF*

QueXF, a CADE (Computer Assisted Data Entry) Tool, processes paper forms that were created in queXML, such as survey questionnaires. It reduces error and fatigue by eliminating manual data entry. OMR (Optical Mark Recognition) is performed on each form to determine if boxes have been filled. On being trained, QueXF makes use of ICR (Intelligent Character Recognition) to detect handwritten characters. A verifier (using a web browser) will then confirm that the entered fields are correct.

3. PROPOSED METHODOLOGY

I) Working of the Target Application

Though we are building a generalized system for digitization of paper forms, for the initial implementation we have chosen to focus on a single target application. This target application extracts data from the student admission forms of CBSE schools. This board has largest affiliated schools in India. The school office distributes admission forms at starting of academic session every year. Every form has two sides; each side is filled by rows of information that the student fills their personal information. The contact no, Aadhaar No, Samagra ID etc.

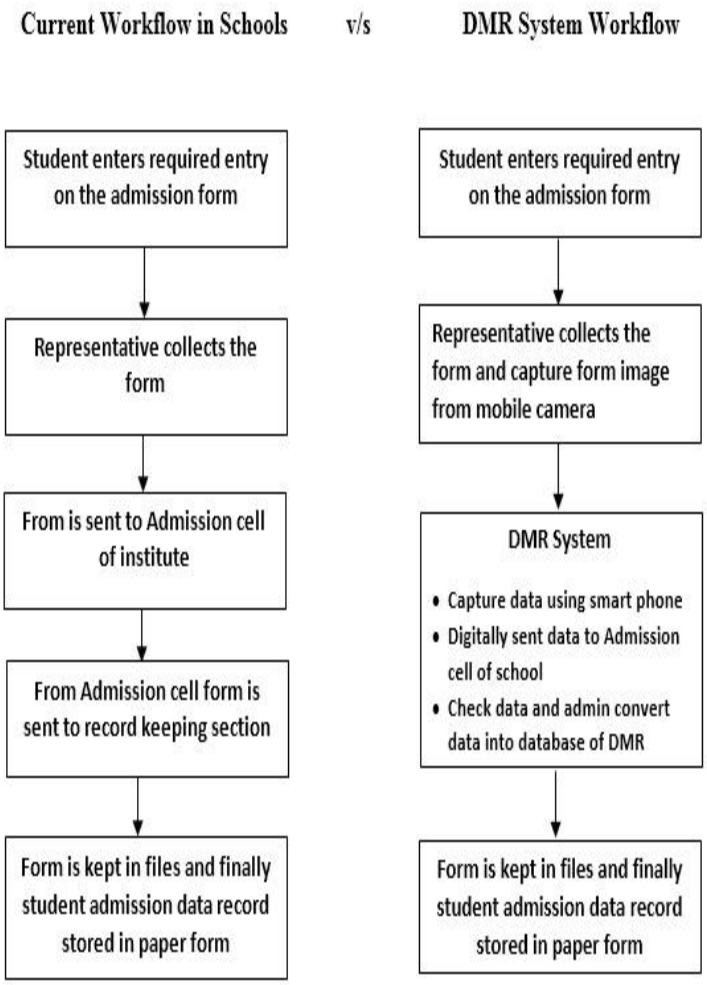


Figure 1: Comparison of Work Flow v/s DMR System

Figure 1 represents two process flows. The admission form is sent to administration department via the reception of the school. The admission cell selects the student form that they require. The administration office accumulates the data from these forms and then derives the total count number student to be admitted. The admission cell officer form contains fields like name, address, contact no, Aadhaar No, Samagra ID etc.

[1] It may need to double check the accumulated data as it is captured manually and this takes up additional time. In this manner the required data is also maintained by the administration office. However, in the proposed system – DMR.

[2] Once the student fills his/her admission form, the representative of reception captures the data from the filled form using smartphone camera. This data is communicated digitally to the administration office, which redirects to the concern authority of school data record management. Finally, this data is to be used in creation of database file of DMR application. This process saves bulk data entry and management cost associated with the current process.

II) DESIGN ASPECTS

Further analysis of the problem brings to notice several design considerations. DMR does not require that the paper forms to be digitized be particularly designed to work with the application. This is important because organizations often find it difficult to change or redesign form templates that are already in use. This current design makes sure that the form need not be altered to fit the requisites of the application.

III) IMPLEMENTATION

Following are the main processing steps needed to digitize form data:

The first step in the process flow of DMR is to use a smartphone's camera to capture an image of the empty form to be used as a template. The specifications of the camera affect the quality and thereby also the results. The user needs to capture the photo with steady hands keeping in mind that the images should be well focused. It should include all of the form contents and the background should be minimal. Background and other irrelevant details add to the noise and tend to reduce the accuracy of the system. Note that: the further

stages of implementation make use of algorithms provided by Open CV, an open source computer vision and machine learning software library. This library contains several algorithms that can be used to process images in real time over the mobile platform.

A. Template Registration

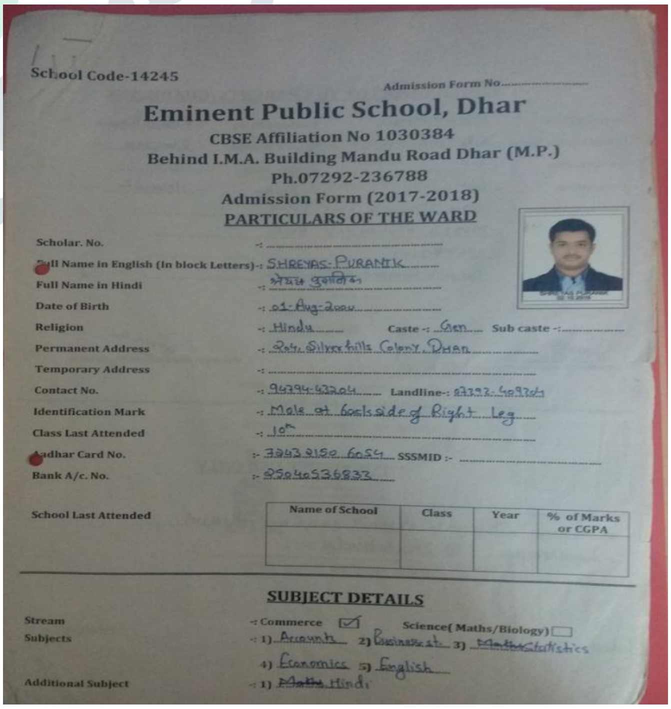
Templates along with their unique features are stored in a database. Feature detection algorithm is applied to compute gradients based parameters at every pixel, to verify if it can serve as a good feature point for the image. This process aims at finding high levels of curvature in the image gradient. A group of such feature points serve as a strong identifier for the image. For each feature point mathematical statistics like gradient magnitude and angle are computed using feature description algorithms. These mathematical descriptions of key points are required in matching and mapping stages. DMR uses the FAST algorithm for feature point detection and BRIEF algorithm for feature point description. FAST is robust, simple to implement and is faster than various feature detector algorithms. Alternatively ORB feature detector and descriptor can also be used. ORB is a one of the fastest binary descriptor based on BRIEF. It is independent of rotation and resistant to noise.

Feature Detector Algorithm	Feature Descriptor Algorithms	Sample	Key points	Execution Time
FAST detector	Brief Descriptor	formFront1	22442	0.287301
		formFront2	23065	0.285611
	ORB Descriptor	formFront1	22442	13.203804
		formFront2	23065	13.351138
ORB detector	Brief Descriptor	formFront1	500	1.050528
		formFront2	500	1.0583
	ORB Descriptor	formFront1	500	1.908036
		formFront2	500	1.996605

Table 1: Comparison of various detection algorithms

B. Data Field Mark-up

In this stage, the user is expected to mark regions of interest over the template. For e.g.: In an admission form if the user is interested in the Aadhaar Number field, he will use the interface to mark and label the region where Aadhaar Number field is shown. OCR will then be applied on the quantity field to extract the required information.


School Code-14245
Admission Form No.
Eminent Public School, Dhar
CBSE Affiliation No 1030384
Behind I.M.A. Building Mandu Road Dhar (M.P.)
Ph.07292-236788
Admission Form (2017-2018)
PARTICULARS OF THE WARD
Scholar. No. -:
Full Name in English (In block Letters) -: SHREYAS PURANIK
Full Name in Hindi -:
Date of Birth -: 01-Aug-2000
Religion -: Hindu Caste -: Gen Sub caste -:
Permanent Address -: 204, Silver hills Colony Dhar
Temporary Address -:
Contact No. -: Landline-:
Identification Mark -: Mole at back side of Right Leg
Class Last Attended -: 10th
Aadhar Card No. :- SSSMID :-
Bank A/c. No. :-
School Last Attended

Name of School	Class	Year	% of Marks or CGPA

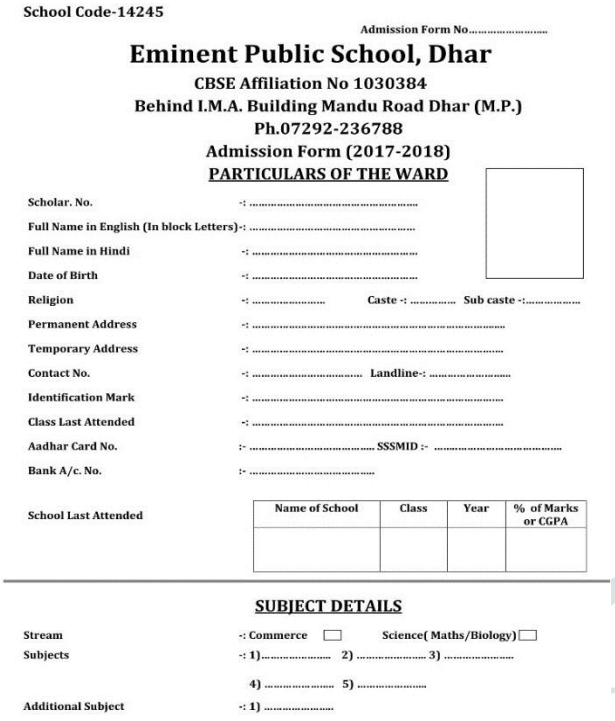
SUBJECT DETAILS
Stream -: Commerce Science(Maths/Biology)
Subjects -: 1) Accounts 2) Business st. 3) Statistics 4) Economics 5) English
Additional Subject -: 1) Hindi

C. Matching

This stage involves making spatial changes to the image of the filled form to align with the template that is stored on the application. The stage of aligning is necessary to ensure thatthe entire form has been taken in. We apply feature detection and description algorithms on filled form. This data is then compared with the template data in database, to classify the filled form and find its corresponding template. This comparison is done using the FLANN algorithm. The FLANN is a Neural Network based algorithm that makes use of nearest neighbors to select matching key points .

D. Mapping

Once appropriate template is identified, regions of interest previously marked on the template during Data Markup Stage need to be mapped on the corresponding area of the filled form. These regions are first loaded according to the template identified. Then these markups are mapped on the input image using RANSAC algorithm



School Code-14245

Admission Form No.....

Eminent Public School, Dhar

CBSE Affiliation No 1030384
Behind I.M.A. Building Mandu Road Dhar (M.P.)
Ph.07292-236788
Admission Form (2017-2018)
PARTICULARS OF THE WARD

Scholar. No. -: .....
Full Name in English (In block Letters)-: .....
Full Name in Hindi -: .....
Date of Birth -: .....
Religion -: ..... Caste -: ..... Sub caste -:.....
Permanent Address -: .....
Temporary Address -: .....
Contact No. -: ..... Landline-: .....
Identification Mark -: .....
Class Last Attended -: .....
Aadhar Card No. :- ..... SSSMID :- .....
Bank A/c. No. :- .....

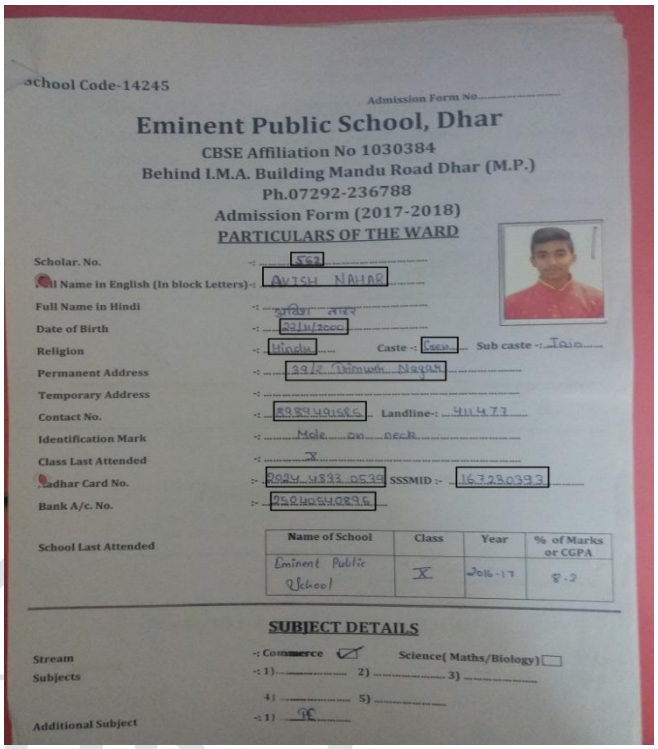
School Last Attended

Name of School	Class	Year	% of Marks or CGPA

SUBJECT DETAILS

Stream -: Commerce ☐ Science(Maths/Biology) ☐
Subjects -: 1)..... 2) ..... 3) .....
4) ..... 5) .....
Additional Subject -: 1) .....

Figure 2: Template of Admission Form Figure



School Code-14245

Admission Form No.....

Eminent Public School, Dhar

CBSE Affiliation No 1030384
Behind I.M.A. Building Mandu Road Dhar (M.P.)
Ph.07292-236788
Admission Form (2017-2018)
PARTICULARS OF THE WARD

Scholar. No. -: 553
Full Name in English (In block Letters)-: AVISH NAHAR
Full Name in Hindi -: अविश नाहर
Date of Birth -: 23/1/2000
Religion -: Hindu Caste -: Gen. Sub caste -: Jain
Permanent Address -: 39/2 Dalmiya Nagar
Temporary Address -:
Contact No. -: 8989491565 Landline-: 411477
Identification Mark -: Mole on neck
Class Last Attended -: X
Aadhar Card No. :- 2024 4833 0539 SSSMID :- 16723093
Bank A/c. No. :- 25040540896

School Last Attended

Name of School	Class	Year	% of Marks or CGPA
Eminent Public School	X	2016-17	82

SUBJECT DETAILS

Stream -: Commerce ☑ Science(Maths/Biology) ☐
Subjects -: 1) ..... 2) ..... 3) .....
4) ..... 5) .....
Additional Subject -: 1) IT

Figure 3: Demo of key point Detection on the form

E. Optical Character Recognition

For recognizing characters in the mapped regions we make use of Tesseract OCR by Google. Tesseract is the most accurate open source OCR engine available. Combined with the Leptonica Image Processing Library it can read a wide variety of image formats and convert them to text in over 60 languages.

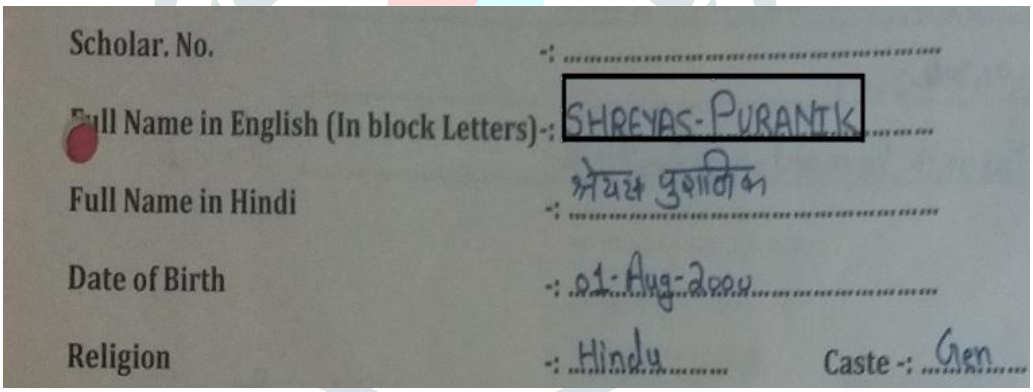


Figure 4: Marking sample for OCR

4. CONCLUSION

This paper presents a basic implementation of DMR which mainly focuses on recognizing the sections of the filled form image which are of value to the user. Numerous changes and refinements have to be made in order to achieve higher accuracy in recognizing text from the image. Development of a more robust handwriting recognition system will complement the usability of DMR. The preliminary test conducted deals mainly with the accuracy of form identification and character recognition. More extensive testing needs to be done with varying lighting conditions and out of focus images.

In some extreme cases, forms with very high number of fields give poor performance in form identification stage. To assist the form identification stage, detection and identification of logos and other areas in form that will uniquely identify the form can be considered. This will greatly improve the performance of the system.

Finally, though DMR is designed to output data in CSV format the export interface can be extended to align with other established formats so it can be easily integrated with other existing applications on the PC or via web. This data that is extracted can be used to populate a schema less database which is then queried on the fly to get insights on the collected sections.

5. REFERENCES

[1] D. Doermann, J. Liang, and H. Li, "Progress in camera-based document image analysis," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pp. 606–616, IEEE, 2003.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol.60, no. 2, pp. 91–110, 2004.

[3] T. Parikh, P. Javid, S. K, K. Dhosh, and K.Toyama. Mobile Phones and Paper Documents: Evaluating a New Approach for Capturing Micro Finance Data in Rural India. In CHI 2006.

[4] 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.

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

6. AUTHOR BIOGRAPHY

GAURAV GUPTA was born in Amjhera Dhar, India, in 21/05/1987. He received the MCA degree in computer application from the Mahakal College Ujjain affiliated to Rajiv Gandhi Technical University, Bhopal M.P. India, in 2010, also He completed B.Ed. degrees in 2015. & joined the Department of Computer Application and Science, in LSA College Dhar, M.P. as an Assistant Professor. Since September 2011, he has been with the PGT computer science in Eminent Public School Dhar M.P. His current research interests include image processing, E-governance. Currently Gupta is a research Scholar of the Dr. A.P.J. Abdul Kalam University in Indore M.P India.