OCR Generic Testing Strategy[DRAFT]

Version 1.0

Funded by

Technology Development for Indian Languages DeitY

Revision History

Sr. No	Version	Date	Created by	Comments
1	1.0	17 Nov 2015	CDAC GIST Pune	Version 1.0 for Generic OCR testing Strategy

Table of Contents

CHAPTER I: Testing Methodologies

1. INTRODUCTION

Optical character recognition (OCR) is the conversion of images into machine-encoded text. Optical Character Recognition has been an active subject of research since decades. Despite the age of the subject, it remains one of the most challenging and exciting areas of research in computer science. Software testing and evaluation is the integral part of the development cycle of any software product Therefore, development of these technologies/products requires close monitoring through testing and evaluation at various phases.

2. PURPOSE

The purpose of this document is to propose a Testing & Evaluation strategy for the OCR system, which can be referred by any testing agency for determining the overall quality of a OCR system.

3. SCOPE

The scope of this document is to define OCR system testing strategy which mainly includes importance of test data and various testing techniques.

Document scope is generic in nature taking into account the future development of OCR system. However, testing will be done as per the present limitation and specification of the system.

4. WHY TESTING IS IMPORTANT?

Software testing is an important phase in software development life cycle as it verifies and validates the system under test i.e. whether it works as expected and satisfies the stakeholders need.

With respect to OCR system also, testing & evaluation is significant; as it is important to test the system before deployment.

In order to assess the system output, appropriate quality assessment techniques should be adopted for determining the system performance in comparison to the benchmark level or with the quality of previous version or with similar kind of different product.

5. TESTING PROCESS

5.1 Evaluation Parameters of OCR system

We believe, though this project the perspective on the performance of the OCR systems would be made by illustrating and explaining the actual OCR errors to their finer details. The analysis made through this angle would provide the insights into strengths and weaknesses of OCR systems under development and perhaps provide the road maps to future progress. While doing this, the reports generated would also furnish pointers to improving accuracies at various stages in development. Throughout this project our focus will remain on examining in some detail the errors made by OCR systems under development for Indian languages and to speculate about their cause and possible remedy.

Throughout the testing lifecycle of this project the OCR systems will be evaluated for the effects of various scanner settings which includes different image formats, resolution, image type , e.t.c. Character level accuracy (on books, magazines, newspapers), word level accuracy (on books), observation for the outcome of layout retention module. Some of the major evaluation parameters are given below, but they may undergo changes over the course of testing lifecycle.

5.2 Calculating accuracy

Each character/Akshara insertion, deletion or substitution needed to correct the OCR generated text can be considered as an error. Each reject symbol/marker is also counted as substitution error in this calculation. All the characters are represented in Unicode format. Character accuracy can be defined as

$$\boxed{\%\,Character\ Accuracy\ =\ \left(\frac{N_{truth} - N_{error}}{N_{truth}}\right) * 100}$$

Where, N_{truth} is the total number of characters in ground truth text.

The above standard string edit distance gives possible errors related to single character insertion, deletion and substitution. In case the number of insertion, deletion, substitution errors N_{error} grows more than the N_{truth} then the accuracy obtained by above equation is Negative. Such negative figures are marked as <0 wherever possible. But such negative errors render very little relevance of OCR to end user.

$$\boxed{\%\, Word\ Accuracy\ = \left(\frac{N_{truth} - N_{werr}}{N_{truth}}\right) * 100}$$

The word level accuracy is calculated by measuring total aligned words in OCR output which are correctly recognized to the actual words in ground truth.

5.3 Speed of OCR

The speed of the OCR system is represented in ACPS. It also should consider the number of accurate characters per seconds. Hence speed of OCR systems can be calculated as,

$$\boxed{Speed\ (ACPS) = \frac{Number\ of\ Accurate\ Characters\ in\ Output}{Time\ (Seconds)}}$$

$$\boxed{\text{Accurate characters in O/P} = \text{Total no. of char. in ground truth} * \text{ Character Level Accuracy}}$$

5.4 Evaluation of Complexity in Printed Scripts

Indian languages are complex in nature especially ligatures in some scripts and font/glyph variations. The complexity in scripts is apparent in the printed format with variations in horizontal and vertical style of representing conjuncts for some scripts. The accuracy of OCR for recognizing such diversely complex representations is calculated differently.

Similarly the printing technology has evolved over the period of time presenting the challenges with the contents printed with different technology.

5.5 User Level Testing

The evaluation metrics mentioned above classify different OCR errors and provides information from more core-engines point of view. The user level testing involves testing of following specifications,

- Testing of application for proper UNICODE support.
- Test for Image Input Mode (Scanner/Image-File).
- Test for Input Image Support (Color / Grey/ Black-White)
- Supports for File Input Format (TIFF/BMP/JPG/PNG)
- Supports for File Output Format (RTF/DOC/ODT/TXT)
- Application mode (Web/Desktop (Windows/Linux) based)

CHAPTER – II : Test Data

6. Selection of Test Data

To objectively measure progress in character recognition technology and to identify research problems, two types of evaluation are needed: internal evaluation and independent evaluation. In internal evaluation, researchers own test data sets or standard (public) test databases are used to measure and compare their progress. The creation and distribution of a variety of standard test databases is an important task in the OCR research. The determination of the fitness of a system for a purpose---will it do what is required, how well, at what cost, etc. Typically for a prospective user, it may be comparative or not. Such adequacy evaluation is also targeted in this testing and test data is selected from both developers and end-users angles.

6.1 Data creation

System accepts image as input, following parameters can be used for image creations

6.1.1 Device Type

1. Scanner
   a. Flat bed
   b. Sheet feed
   c. Drum Scanners
   d. Portable scanners
2. Camera captured images/Scene Text

3. Fax Images

6.1.2 Resolution

Images are scanned in different resolutions. The resolution depends on the type of scanner being use. It varies from,

1. 75 DPI
2. 150 DPI
3. 300 DPI
4. 600 DPI
5. 900 DPI

6.1.3 Page Quality Parameters

Since image creation is manual process, it poses its own set of challenges,

1. Skew
2. Noise
3. Paper quality
4. Camera angle
5. Conditions during capturing photo
6. Distance from source

DATA Corpus creation contains many variations:

6.2 Data Sources

As name suggests the data from various sources like Books, Magazine, Newspaper should be identified.

While identifying following parameters should be considered,

1. Paper quality
2. Paper size
3. Print size
4. Font size
5. Publication date
6. Publication House
7. Printed by
8. Printing technology used

Also variation in data like

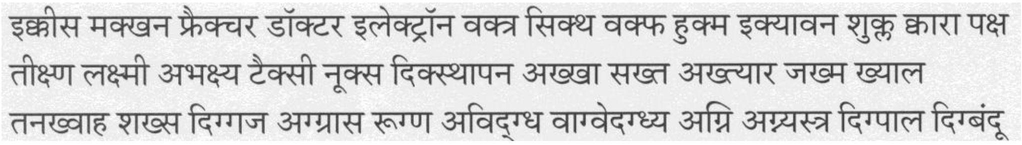
1. Single column
2. Multicolumn
3. Single font
4. Multi font
5. Single font size

6. Multi font size
7. Single language data
8. Multilingual data
9. Special characters
10. Decorative font

6.3 Synthetic Data

Basic Characters – Since basic characters are building blocks of any language, the 1st level of data consists of basic characters printed in various fonts, font sizes & printers

Syllables/Conjuncts – Then simple syllables (consonant plus matras) & word dataset having complex conjunct combinations can be prepared, as shown below. This set should contain all possible combinations of syllables.



इक्कीस मक्खन फ्रैक्चर डॉक्टर इलेक्ट्रॉन वक्त्र सिक्थ वक्फ हुक्म इक्यावन शुक्क क्कारा पक्ष
तीक्ष्ण लक्ष्मी अभक्ष्य टैक्सी नूक्स दिक्स्थापन अख्खा सख्त अख्त्यार जख्म ख्याल
तनख्वाह शख्स दिग्गज अग्ग्रास रूग्ण अविद्ग्ध वाग्वेदग्ध्य अग्नि अग्न्यस्त्र दिग्पाल दिग्बंदू

Fig: Snippet of Hindi Conjuncts

Randomly picked paragraphs - The random data is to be chosen from the websites. Paragraphs are to be chosen for each script with addition of font variations and size variations. The paragraphs are to be randomly taken from websites like www.bbcworld.com , www.wikipedia.org etc. The data is first to be cleaned and checked for any in-accuracies.

6.3.1 Procedure of Creation of Synthetic Corpus

- The data required for testing includes randomly selected paragraphs from web, Consortia annotated corpus, Conjunct variations.
- This data is first to be created for each script under test.
- The generated data is to be verified from different experts.
- Now quipped with data, the take printouts of the same data on standard LaserJet printers with following possible variations
    - Font
    - Font Size
    - Mixed fonts
    - Mixed font sizes
- Scan printouts at different resolutions (200dpi, 300dpi etc) without introducing any skew.
- Proper scanning instructions should be conveyed to the member responsible for scanning.
- All this procedure should be carried out by skilled persons so as to avoid any problems.

6.3.2 Criteria for Font Style Variations

Many different varieties of fonts and typefaces are now being available for all Indian languages; testing for all these fonts poses a challenge. There is dearth of proper survey or availability of fonts and their faces for every Indian script. But with the release of software CD’s in every languages by TDIL there is a common database of easily available and possibly popular fonts. Such popularly known fonts for different languages can be used.

6.3.3 Point Sizes chosen

Since the data is synthetically generated using a computer, it allows good mix of font size variations unlike the fonts found in older books. The font sizes can be chosen after analyzing font-style in question and also the script. Since Indian scripts contain more complex formation of glyphs, it requires bigger font sizes for legibility unlike their English counterpart.