A Framework for Compilation of Multi-Lingual Handwritten Database: Four Levels XML Ground-Truth

Prakash Choudhary*, Neeta Nain† and Manindra Nehra†
*Department of Computer Science and Engineering
National Institute of Technology Manipur
Imphal, India–795001
Email: choudharyprakash@nitmanipur.ac.in
†Department of Computer Science and Engineering
Malaviya National Institute of Technology Jaipur
Rajasthan, India-302017
Email: neetunain@yahoo.com

Abstract—In this paper, we are presenting a semi-automatic framework for annotating multi-lingual handwritten texts document images. There is a significant need for a structure that can annotate the coordinate segmentation information of the text present in a handwritten document image to provide a platform for OCR algorithm evaluation. In this paper, we describe an XML based four level annotations of handwritten text image that contain the ground-truth information of script text image in Unicode format. In order to collect the huge amount of data for linguistic researchers, structure provide a way to store and annotate at different four levels: Image, Lines, Words and Characters which aids for benchmarking of various OCRs. Structure would be best source for compilation of an annotated handwritten corpora in systematic and scientific way by storing a labelling(markup) information of image script texts in a Unicode and an XML file format that encapsulates the bounding box pixel information of each level in a collaborative manner. The structure provides useful results based on the annotation for various qualitative and statistical corpus approaches to linguistic analysis.

Keywords—Document Image Analysis; XML annotation; Ground Truth; Database;

I. INTRODUCTION

There is a significant need for linguistic resources that can be able to train a system for extract and recognize text efficiently from handwritten documents. Annotation of textual document is time consuming and error prone procedure that requires the utmost care. Annotation is the process that make database computer process able and significant for the various linguistic related research such as the grouping of information, indexing and managing a large amount of natural handwritten information. The ground truth database is the basic need in the field of document image analysis because, various parameter are associated with the term ground-truth that provide a standard platform for training and testing of various optical characters recognition techniques.

The handwritten document is sharply different from printed documents, and it’s vary with person to person based on handwriting style. In a country like India, languages vary from state to state, it is common here to interspersed Indic-Script text with the English words in a daily based human writing. A special treatment is required to treat with these multiple scripts. Most of the document either related to the private or public organization are bilingual or multilingual. Thus, an annotation of such multi-script documents images makes them possible for searching and accessing desired information from these documents using the conventional textual search.

IAM [1][2] is the first available corpus for English language, database consists the 1,539 handwritten text images of full length sentences. IAM handwritten documents database were annotated for lines, words in an XML file format manually. PBOK [3] is a collection of 707 text pages written in four scripts; the database is ground-truth for content and pixels information. [4] design a model based approach for annotated online handwritten database. An early attempt at XML tagging approach was used in [5]. [6][7] proposed XML standard annotation tools for annotate Indic scripts online handwritten database.

PixLaber[8], GTLC(Ground-Truthing Text Lines and Characters tool)[9], TRUEVIZ[10] and MAST[11] are some existing structure for annotate offline handwritten database. [12] designed an ILT truthing tool for ground-truth of multilingual Indic script documents.

In this paper, we described a Semi-Automatic framework for annotation of handwritten text images that automatically associate the ground-truth information of handwritten multi-script images. Structure align the plain transcripts text of handwritten data automatically to generate the annotation at the four levels: Image, Lines, Words, and Character. Each level will get a Unique Identification Number(UID) which is the hyphen extension of original image Unique Identification number. As a result, an XML file contains the typical hierarchical representation of handwritten text image meta information which is comprised of region pixel information of a text in the image. Each level of hierarchy contains a mark-up information that captures annotation information at that level. In the end, structure generates the quantitative results of the corpus for various statistical approaches to linguistic analysis.

The rest of the paper is organized as, in section 2, we explain the overview of the structure where different steps involve in the Corpus design methodology: method-
ology for data preparation, process of annotation and brief
description about the graphical user interface. Section
3, describe the automatic generation of an XML meta-
information file for ground-truth. Section 4, A comparative
analysis of CALAM with the existing annotation structure.
At last in section 5 conclusions are presented.

II. OVERVIEW OF THE STRUCTURE

The collection of offline handwritten data for corpus
is started with raw handwritten text images. But to make
these images available for various computer processing
 techniques structure start processing with raw handwritten
image and provide an annotated database that contains
output an image, corresponding XML file of Unicode
transcription and script region bounding box information.

For better classification of data storage and consistency,
we are indexing the ground truth information of the
image in database with auto-generate Unique Identification
Number for each level. The structure follows the same
above top-down process for insert and annotates a new
image.

A. Data Preparation

The structure starts an annotation process with a raw
image. In our experimental work we have chosen two
types of text images sample. In the first example we
have taken images based on daily life handwritten form
as shown in Figure 1 where sample images contain a
multi-script text. In the country like India language and its
writing style is vary from state to state. People are mixing
some words of English language with their native language
words in their daily natural handwriting. As shown in
Figure 2, we have chosen some selective sample of images
where English words are mix with the Indian script Hindi,
Urdu, Meetei and Telugu.

In the second type of input, we have selected a sample
of images from the handwritten corpus. Figure 2(a) shows
the sample image taken from the offline handwritten De-
vanagari database and Figure 2(b) shows a sample image
taken from Urdu handwritten Corpus CALAM [13].

For the maximum variance among the words, the col-
clection of words were divided in categories like News,
Science Technology, History, Literature, Politics, Archi-
tecture and Economics. All the categories, sub-categories
and their notation are as shown in Table I.

The corpus data collection form is designed in a specific
way to make it is convenient, apparent and comfortable
for data collection. Sample layout of the form is designed
on A4 size paper as shown in the Figure 2. The form
is divided into four parts with the description as shown
below:

- The top block header is designed to collect the
  writers information likes name, profession and tick
  on rectangles for gender, agegroup and qualification..
- After header a printed text block designed, which
  consists the printed texts lines of 70 to 80 words.
- Third part having a blank space for replicating printed
text in natural handwriting, in which writers are asked
to write the printed text in their own handwriting.
- At bottom footer block, in which space are available
  for address, signed of writers and unique id of form
  are given.

For syntactic variations, the form were filled by writ-
iers of different age groups, profession and educational
qualifications and from distant locations like shopping
malls, railway stations, bus stand, hospitals etc, because
the handwriting of a persons are gets affected by the mood,
situation and surroundings.

The writers were offered multicolored (red, blue, green
and black) ink and gel pens. Besides this the persons from
different graphical location of India are also involved in
data collection; those who are comfortable with Hindi and
those whose mother tongue is not Hindi.
B. Process of Annotation

In our work, we consider a higher level sentences based handwritten documents that combine a different style of writing in single trial instead of isolated characters or words. Annotation process starts after the collection of all handwritten images. The images were scanned at high resolution 600 dpi and saved in PNG format. We tried to annotate each scanned handwritten document at four levels image, lines, words and characters that are highly prefer platform for providing full support to OCR segmentation as shown in Figure 3[14].

The process of ground truth and annotation for handwritten text corpora starts with the raw text document image. Structure divide image into four level for ground truth, each level get a UID number.

To maintain the consistency throughout the database and annotation process each level get a Unique Identification Number which is auto-generate during the annotation process for corresponding level and indexed the associated information of the image in database with this number. The process of auto-generate 16 bit UID is comprised of language, category, subcategories and number. This will add to classification of desired information according to language or nature of the text.

To compute a Ground Truth (GT) for a handwritten text image page first, we applied image prepossessing for removing noise and correct skew from each text page. After image prepossessing we upload an image and manually draw a bounding box around the first text line present in the text page, and then it will segment and stored image of the first line in the database. Structure fetch and update the all corresponding pixels coordinates belonging to this region automatically in the database and assign a UID for this line. The same process we apply for the first word of line and structure fetch and update bounding box pixels of associated word region in the database with a UID number for word and stored segmented word image. The same process will be applicable till the segmentation of last line and last character and so on. Transcript texts of the text page, each line and words are fetched during the annotation process and display in the side panel.

As a result, we create a database using this structure where all information stored in database and image of text page, segmented lines and words stored separately in system with the corresponding UID as a name of image in PNG format.
C. Structure Graphical User Interface

The structure provide a user friendly interface that given the option to create an annotated corpus for various languages, searching and retrieval of desired information with different criteria based on interest such as: age, education, region, text categories etc.

The structure follows the same progressive gradually flow of annotation process as shown in Figure 3. The user interface is simple to upload and annotate a new document easily. During the insertion and exploring, it displays the image and corresponding annotating information in a collaborative manner at the same time on the same viewport as shown in Figure 5. For the proper visualization of the region, a bounding box appears on the corresponding line in an image. The image and relevant information are on same viewport make it useful for the validation of context information and visual review of annotated data.

The purposed Devanagari handwritten text images corpus are designing by collecting approximately 150 handwritten text images for each categories, with the help of writers of different profile and location. The interface with sample uploaded form is shown in the Figure 5. The annotation of segmented line and word are shown in Figure 5 and Figure 6 respectively. In annotation of the Devanagari handwritten text line and word are shown in the left and their ground truth information in right.

The aim of the interface is to provide a platform for different document analysis tasks. The simplest task of GUI is to explore and create an annotate dataset, markup the regions of scanned documents image as text and assign labels to regions. Based on the assign labels researchers can extract the desired/interested text from the document or directly can retrieve step by step from the hierarchy level of the XML file.

III. GROUND-TRUTH META- INFORMATION SPECIFICATION

Although, the heart of database is offline handwritten text image, the final output received from the structure is a collection of segmented image, line, word and XML encoded meta-information file. The auto-generated output XML file structure layout encapsulated meta-information of the corresponding scanned handwritten image in hierarchical four levels. There was some prior work related to online Indic script annotation[3][9][4].

As a result, structure generates XML file for text page based on entries. The XML files contain the same material as we provided during the data entries in hierarchical four levels granularity meta-information description. A relevant information as per required functioning can be extract from the XML file or directly using the searching and filtering option from the database.
Each level contains a digital transcription of handwritten text in Unicode format and annotation information of bounding box. The level-3 is store some additional grammatical information apart from textual region pixel information such as: antonyms, synonyms, tag and starting letter of the word.

IV. COMPARATIVE ANALYSIS OF THE PROPOSED FRAMEWORK CALAM WITH EXISTING STRUCTURE

A comparatively analysis of CALAM with existing structures Pix Labeler[12], GTLC[14], MAST[8] and TRUEVIZ[7] for offline handwritten text images database shows the functionality of the existing tools of annotation. PixLabler Tools provides a way to annotate English language document. GTLC is available for offline Chinese script handwritten document annotation. Most of the structures focused on either printed documents or isolated words instead of sentences. MAST are designed for annotation of camera based images. Most of the structure is concern about the printed text except the GTLC and APTI.

As compared to the above structures CALAM provides the display of a handwritten text image file and the transcription material of the corresponding image on the same screen in a collaboration context.

CALAM is a simple way for annotation and collection of a large volume of information for handwritten documents, such as: digits, para, lines, words, machine printed text, handwritten text on the same platform. CALAM generate an automatic XML file of annotated meta information that would be useful for ground truth of image (bounding box coordinates of lines, words and ligatures) to benchmarking and evaluation of various OCR technique like segmentation and handwritten text recognition techniques.

The structure uses hierarchical layers of annotation (ground truth information, align transcript texts of segmentation line and word), which is further mapped between database and structure to creates a single XML output which contains multiple layers of information. Besides of NLP domain requirements, structure generates quantitative results of corpus data for various statistical analysis.

V. CONCLUSION

The aim of this work is to annotate a multi-scripts handwritten document and develop a large volume annotate handwritten corpora by the systematic and scientific way along with ground-truth Meta information in XML format. A complete representation of handwritten document image was done in a four hierarchy level: image, line, word, character. The structure provides a progressive process of annotation where each level store the align transcript texts of handwritten data and a bounding box coordinate of the textual region. The digitization format of XML file preserves the documented of the handwritten part and support for computation linguistic and conceptual retrieval of desired information. In addition to raw handwritten image collection corpus. The structure is useful to the parallelism of storage in a consistency manner with the appropriate directory structure and UID file naming conventions. We present a framework for generating a systematic model for benchmarking of various segmentation and OCR experimentation by providing a ground truth.

REFERENCES


