2015 11th International Conference on Signal-Image Technology & Internet-Based Systems

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: neetanain@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 quantitative 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.

PixLabler[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-

978-1-4673-9721-6/15 $31.00 © 2015 IEEE
DOI 10.1109/SITIS.2015.100

IEEE computer society

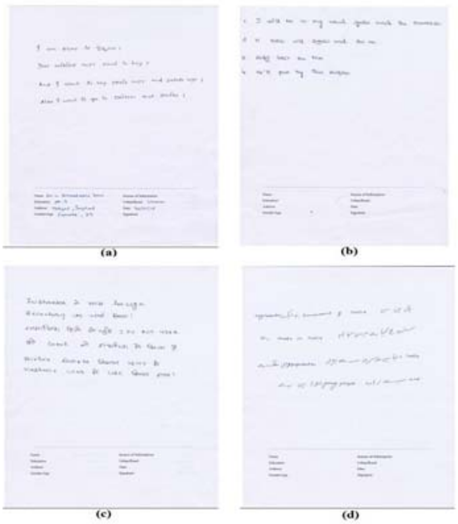


Figure 1. Sample of handwritten form: a) English and Meetei Mayek b) English and Telugu c) English and Hindi d) English and Urdu

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.

Table I
Categorical distribution of data collection in the corpus

Categories	Sub-Categories
News -N	International-IN National -NN Sports -SN
Science or Technology-S	Engineering-ES Medical -MS Physics -PS Chemistry -CS
History -H	Word History-WH Indian History -IH
Literature - L	Poetry -PL Shayari/Ghazal-SL Biography -BL
Politics -P	Central -CP State -SP World -WP
Architecture -A	Rural -RA Urban -UA
Economy/Business-E	Agriculture -AE Industry -IE

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 Devanagari database and Figure 2(b) shows a sample image taken from Urdu handwritten Corpus CALAM [13].

For the maximum variance among the words, the collection of words were divided in categories like News, Science Technology, History, Literature, Politics, Architecture 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 writers 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.

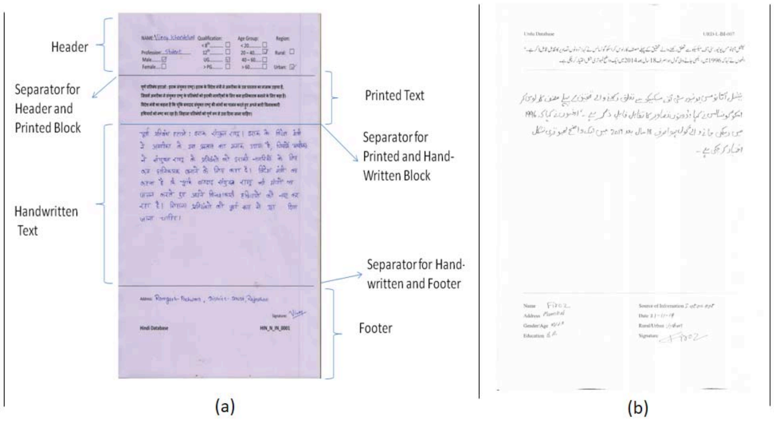

Header
Separator for Header and Printed Block
Handwritten Text
Printed Text
Separator for Printed and Hand-Written Block
Separator for Hand-written and Footer
Footer
(a)
(b)

Figure 2. Sample of Corpus database handwritten form a) Sample of Devanagari handwritten form, b) A Sample image of Urdu handwritten Corpus

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.

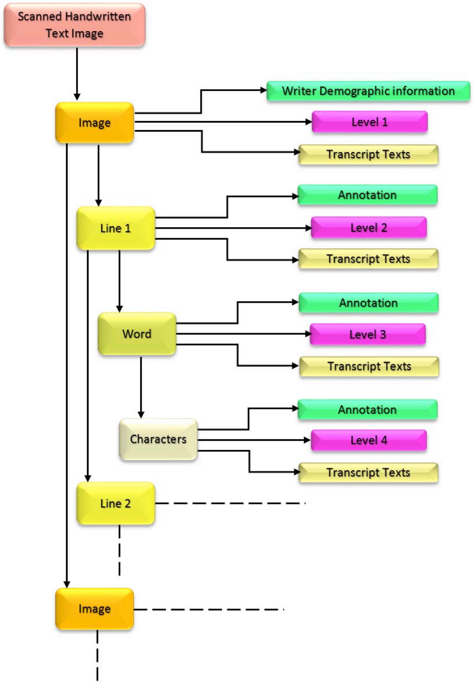
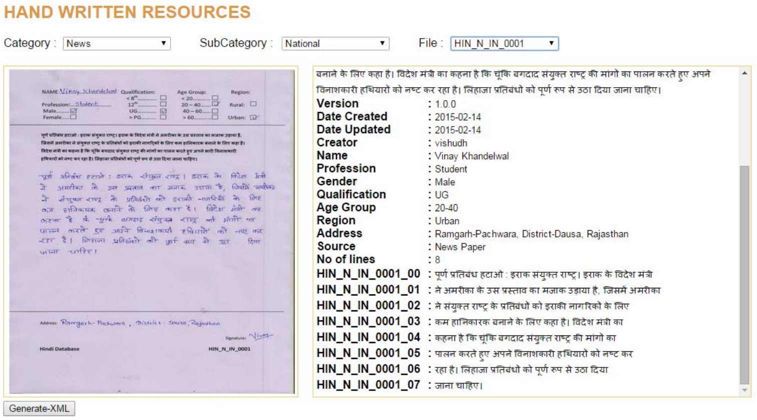


Figure 3. Hierarchical process of four level XML Ground-Truth generation.

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.



Category : News SubCategory : National File : HIN_N_IN_0001

बनाने के लिए कहा है। विदेश मंत्री का कहना है कि चूंकि बगदाद संयुक्त राष्ट्र की मांगों का पालन करते हुए अपने विनाशकारी हथियारों को नष्ट कर रहा है। लिहाजा प्रतिबंधों को पूर्ण रुप से उठा दिया जाना चाहिए।

Version	: 1.0.0
Date Created	: 2015-02-14
Date Updated	: 2015-02-14
Creator	: vishudh
Name	: Vinay Khandelwal
Profession	: Student
Gender	: Male
Qualification	: UG
Age Group	: 20-40
Region	: Urban
Address	: Ramgarh-Pachwara, District-Dausa, Rajasthan
Source	: News Paper
No of lines	: 8
HIN_N_IN_0001_00	: पूर्ण प्रतिबंध हटाओ : इराक संयुक्त राष्ट्र। इराक के विदेश मंत्री
HIN_N_IN_0001_01	: ने अमरीका के उस प्रस्ताव का मजाक उड़ाया है, जिसमें अमरीका
HIN_N_IN_0001_02	: ने संयुक्त राष्ट्र के प्रतिबंधों को इराकी नागरिकों के लिए
HIN_N_IN_0001_03	: कम हानिकारक बनाने के लिए कहा है। विदेश मंत्री का
HIN_N_IN_0001_04	: कहना है कि चूंकि बगदाद संयुक्त राष्ट्र की मांगों का
HIN_N_IN_0001_05	: पालन करते हुए अपने विनाशकारी हथियारों को नष्ट कर
HIN_N_IN_0001_06	: रहा है। लिहाजा प्रतिबंधों को पूर्ण रुप से उठा दिया
HIN_N_IN_0001_07	: जाना चाहिए।

Figure 4. Framework GUI for Ground-Truth of a Devanagari handwritten text page annotation.

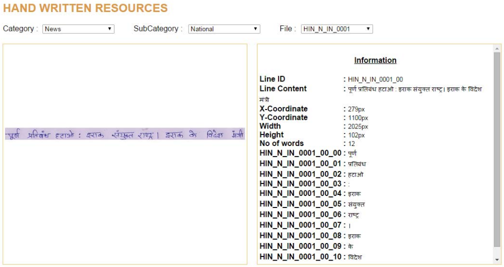
C. Structure Graphical User Interface

The structure provide a user friendly interface that given the option to creation 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.

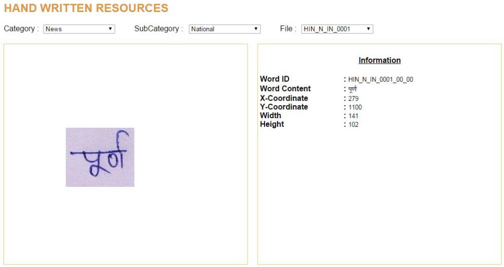


Category : News SubCategory : National File : HIN_N_IN_0001

Information

Line ID	: HIN_N_IN_0001_00
Line Content	: पूर्ण प्रतिबंध हटाओ : इराक संयुक्त राष्ट्र। इराक के विदेश मंत्री
X-Coordinate	: 279px
Y-Coordinate	: 1100px
Width	: 2025px
Height	: 102px
No of words	: 12
HIN_N_IN_0001_00_00	: पूर्ण
HIN_N_IN_0001_00_01	: प्रतिबंध
HIN_N_IN_0001_00_02	: हटाओ
HIN_N_IN_0001_00_03	: :
HIN_N_IN_0001_00_04	: इराक
HIN_N_IN_0001_00_05	: संयुक्त
HIN_N_IN_0001_00_06	: राष्ट्र
HIN_N_IN_0001_00_07	: ।
HIN_N_IN_0001_00_08	: इराक
HIN_N_IN_0001_00_09	: के
HIN_N_IN_0001_00_10	: विदेश

Figure 5. Gorund-truth of a segmented line from the handwritten text-page GUI



Category : News SubCategory : National File : HIN_N_IN_0001

Information

Word ID	: HIN_N_IN_0001_00_00
Word Content	: पूर्ण
X-Coordinate	: 279
Y-Coordinate	: 1100
Width	: 141
Height	: 102

Figure 6. Snapshot of a segmented word annotation GUI

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

[1] U.-V. Marti and H. Bunke. A full english sentence database for off-line handwriting recognition. In Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR 99), pages 705-708, Sept 1999.

[2] U.-V. Marti and H. Bunke. The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition, vol.5(1), pages 39-46, 2002.

[3] A. ALAEI, U. PAL, and P. NAGABHUSHAN. Dataset and ground truth for handwritten text in four different scripts. International Journal of Pattern Recognition and Artificial Intelligence, vol.26(04), pages 001–025, 2012.

[4] A. Kumar, A. Balasubramanian, A. Namboodiri, and C. Jawahar. Model-Based Annotation of Online Handwritten Datasets. In Tenth International Workshop on Frontiers in Handwriting Recognition. Suvisoft, 2006.

[5] M. Agrawal, K. Bali, and L. Vuurpijl. Upx: A new xml representation for annotated datasets of online handwriting data. In Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR 05), pages 1161-1165, IEEE Computer Society, 2005.

[6] S. Belhe, S. Chakravarthy, and A. G. Ramakrishnan. Xml standard for indic online handwritten database. In Proceedings of the International Workshop on Multilingual OCR, pages 1-4, New York, NY, USA, 2009.

[7] U. Bhattacharya, R. Banerjee, S. Baral, R. De, and S. Parui. A semi-automatic annotation scheme for bangla online mixed cursive handwriting samples. In Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 680-685, Sept 2012.

[8] E. Saund, J. Lin, and P. Sarkar. Pixlabeler: User interface for pixel-level labeling of elements in document images. In 10th International Conference on Document Analysis and Recognition, pages 646-650, July 2009.

[9] F. Yin, Q.-F.Wang, and C.-L. Liu. A tool for ground-truthing text lines and characters in off-line handwritten chinese documents. In 10th International Conference on Document Analysis and Recognition, pages 951-955, July 2009.

[10] T. Kanungo, C. H. Lee, J. Czorapinski, and I. Bella. Trueviz: a groundtruth/metadata editing and visualizing toolkit for ocr, 2001.

[11] T. Kasar, D. Kumar, and A. G. Ramakrishnan. Mast: Multiscript annotation toolkit for scenic text. In Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, pages 1-8. ACM, 2011.

[12] V. Govindaraju and R. Vemulapati. Tools for enabling digital access to multi-lingual indic documents. In Proceedings of the First International Workshop on Document Image Analysis for Libraries, pages 122-133. IEEE, 2004.

[13] P. Choudhary, N. Nain, and M. Ahmed. A unified approach for development of urdu corpus for ocr and demographic purpose. In Proceedings of the SPIE 2014 7th International Conference on Machine Vision(ICMV 15), vol. 9445, pages 526-530, 2015.

[14] Prakash Choudhary, Neeta Nain and Mushtaq Ahmed. A Structure for Annotation and Ground-truthing of Urdu Handwritten Text Image Corpus. In Proceedings of the 7th International Conference on Corpus Linguistics (CILC2015), Procedia - Social and Behavioral Sciences, vol. 198, pages 84–88, 24 July 2015.