A Comprehensive Database for Offline Sindhi Handwritten Text Recognition

Author's Details:

Shafique A. Awan[1], Dil Nawaz Hakro[2], Intzar Ali Lashari[4], Zahid Hussain[3], Akhtar H. Jalbani[3],Maryam Hameed[2]

[1]Benazir Bhutto Shaheed University, Lyari Karachi, Sindh, Pakistan

[2]Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan

[3]Department of Information Technology, QUEST, Sindh, Pakistan

[4]Institute of Business Administration, University of Sindh, Jamshoro, Pakistan

Abstract

A well designed image database is very necessary for the recognition of any language and many of language of the world have their own database for the text recognition. In this paper we are presenting the comprehensive database for Sindhi Language which is highly demanding language of the Middle East countries. The database will consist of the many words, which are written by the many writers. This is novel approach for the creating and testing of Sindhi Text. This database contains the isolated characters as well as the text of Sindhi language. The database is consisting of the images of the handwritten text, ground truth information, boxes for the users writing and lines for the text recognition. Tools for the segmentation and creation of the database have been developed in MATLAB, which helps to extract the handwritten data from the designed forms. We developed the different forms for the collection of data from the different users. This database can be used by the other language which possesses the same characteristics of Sindhi Language.

Keywords: *Offline Sindhi Handwritten, Text Recognition*

Introduction

Offline handwriting recognition is still very challenging job which need some perfection at high level. A lot of work has been done by the advanced language such as latin, English, chines, japnes and others which are nearly perfection level but the languages which are cursive in nature need some more attention and perfection. As the Sindhi is cursive in nature and majorly speaking and writing in middle east countries, which possess the number of characteristics of Arabic language. Arabic language cannot be applied to Sindhi language because Sindhi has 52 characters with major four dots which are not included in Arabic language. There is no generally database for the Sindhi handwritten Language yet which can be tested and trained. Sindhi Language is considered as one of the seven language of heaven and has a 5000 years old history and also has a very huge amount of literature (Alana G.A.,1993). There is a absence of databases for Arabic and Arabic related characters. Luqman and others purposed the multiscript database for the testing and training, which can recognize the Arabic language and other related languages. Additionally no multiscript database accessible on a this stage, to test and train the multiscript language (Luqman et al,. 2014).

There is lacking of high level database is available for Arabic language. (AbdelRaouf et al., 2008). Hence different research use the different database of Arabic to test the data which is only helpful to recognize the limited number of words. Some standard database is needed for the recognition of sindhi language and other language database cannot be used for the sindhi language. This database can be used by the different aspect like bank cheques, handwritten library books, postal addresses etc. Here there is no power over the writer composing or composing styles (Somaya et al, 2004). For-example, an arbitrary handwritten words may be produced by a felt pen and could be separate, touching, overlapping characters, half cursive or completely cursive words.(Hull J.J, 1994)

Characteristics of Sindhi Language

Sindhi language is widely spoken by the major areas of Middle East countries especially in Pakistan and India. Sindhi language can be written from the right to left and having 52 characters with different numbers dots position as shown in figure 1. Writing in Sindhi language is cursive in both OCR and ICR. [Sabri A.

Muhammad]. It is very difficult to understand the nature of the character shape because a character changes shape according to its preceding and following position shown in figure 2.

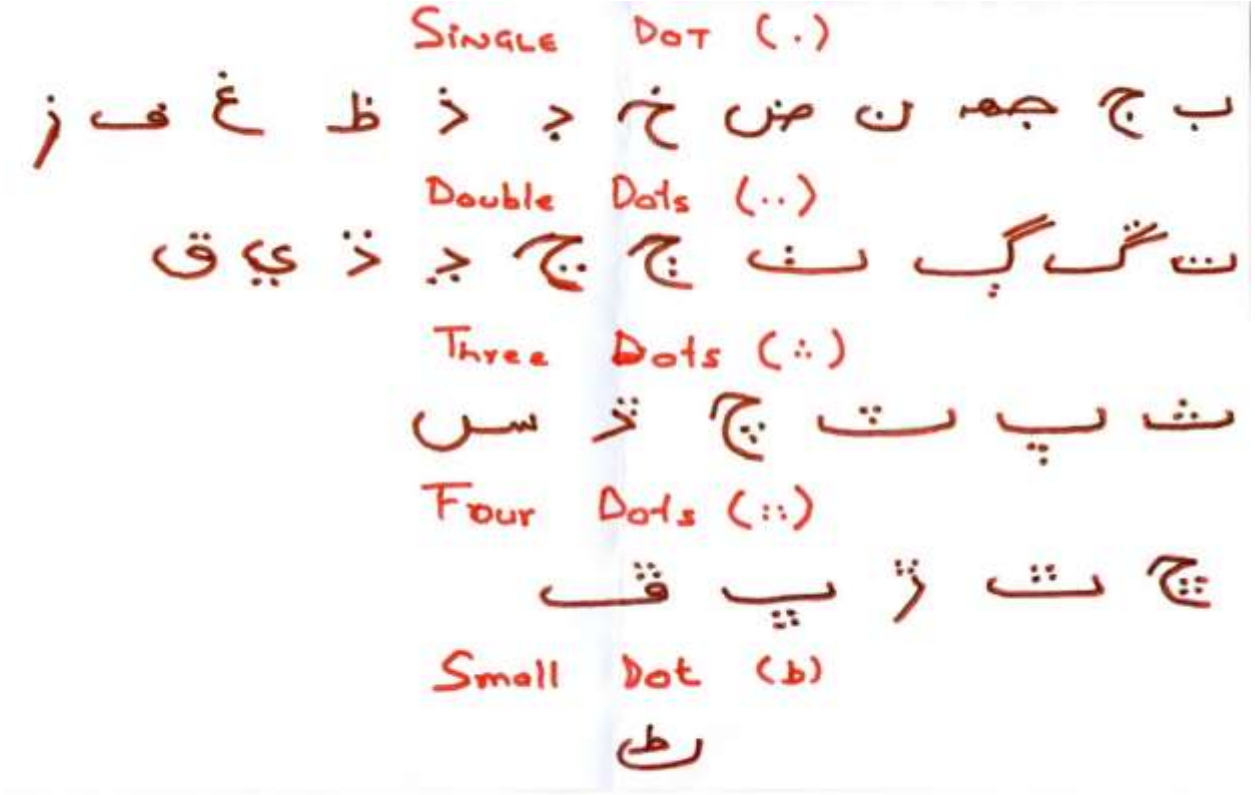


Figure 1: The 52 basic characters in Sindhi Language.

Overview of Recognition System

The general framework of the text recognition is given in figure 7. The basic purpose to create the database is to convert paper document to digital form, usually grayscale or color image data. Recognition system perform the different task like the noise reduction from the images, segmentation of the text from the image and binariztion. Recognition system may be the segmentation free or segmentation based in other words to recognize the limited words or any words respectively. The recognition system also correct the position of text by normalization process where size, slant, skew and lines of text are corrected. Then the feature extraction techniques will be applied which helps to extract the features of image. Core work is the approximation of baseline of each term to the feature extraction more operative.

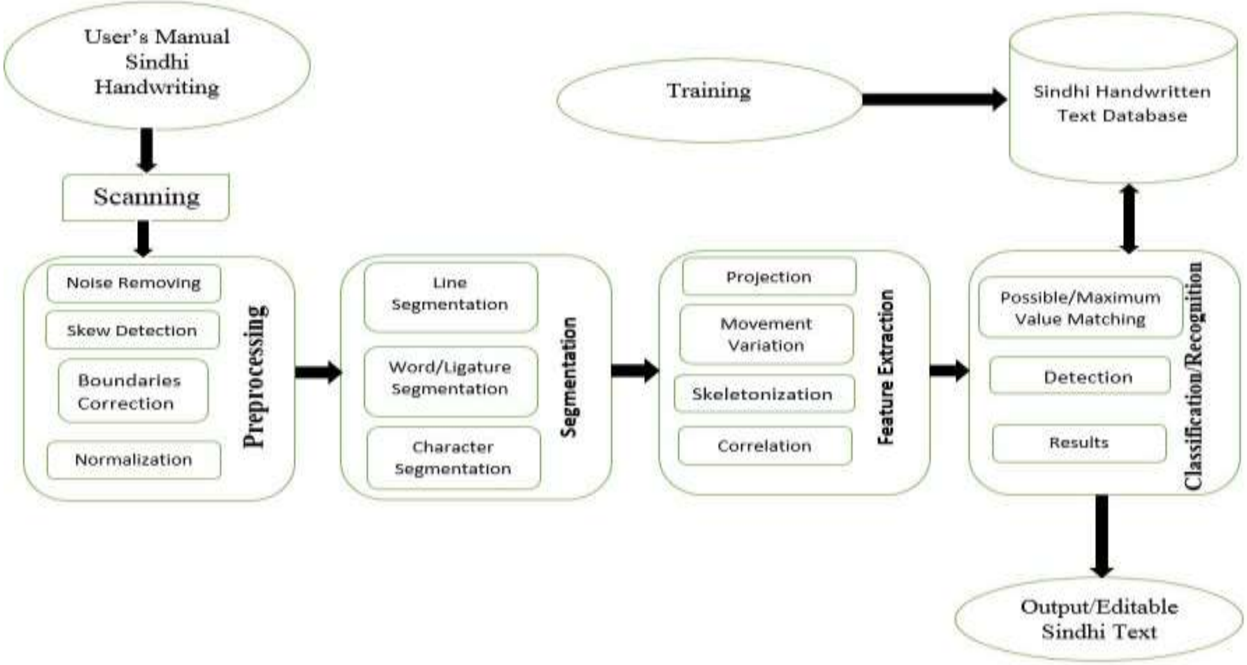


Figure. Complete frame work for the recognition of any Text from the images

Related Work:

There are lot of efforts have been taken for the developing of database for handwritten text recognition and still needs more attention. From the initial, Abhaiba and other developed a database of 2000 words for Arabic language these words were unconstraint in nature and written by four different writers.(Abuhaiba et al, 1994)

In last ten years, a lot work has been done for developing of database for Latin scripts(Al-Badr and Mahmoud., 1995). However no any efforts have been taken yet to develop the database for Sindhi Language.

One of the huge database of IFN/ENIT database was developed by the Institute of Communication and Technology (IFN). Which contains 26,549 images of Tunisian village/town and written by the 411 writers. (Mario Pechwitz et al., 2012)

Another Arabic check database developed in 2003 by the Al-Ohali and others. This database were used by the banks to recognize the checks signature and other data. The database contains the Handwritten Arabic digits and limited numbers of vocabulary text.(Al-Ohali et al., 2003)

El-Sherif and others presented a database for Arabic language which is consist of 70,000 digits written by the 700 writers. He used 60000 arabic digits for the training of database and 10000 for the testing(El-Sherif and Abdelazeem., 2007)

Alamri with other recognize the digits by training the database. Where he tested special symbols and digits by using his database (Alamri et al., 2008). Another database for the indian digit was presented by Mahmoud, where 44 writers wrote the 48 samples of each digit (Mehmoud., 2008)

Ramdan and others created an Arabic Handwritten Database (AHDB) for the Arabic handwritten recognition. The database contains Arabic text images containing open vocabulary. The database contains all forms of the Arabic characters such as first, last, middle and isolated. A total of five scholars having different age group and varying qualification to write the Arabic text line by using the pen of their own choice. A total of 497 images each image contain the town name handwritten along with its ground truth information and scanned at 300 dpi.

The database can be used for handwritten recognition, word spotting and writer identification for Arabic (Ramdan et al., 2013)

Wahab with other produced a Pashto text image database and presented a shape analysis of Pashto characters. The database is comprise of the images are created synthetically and the ligatures are selected from a Pashto novel written by Muhammad Ajan Yar titled "Da Jwand ao Da Ceray". The Pashto database contains images of 1000 ligatures in four font sizes a total of 4000 ligatures of Pashto language. The database is titled as Fast- NU Pashto image database. The database were created at Peshawar University Pakistan(Wahab et al., 2009)

Overview of database Creation

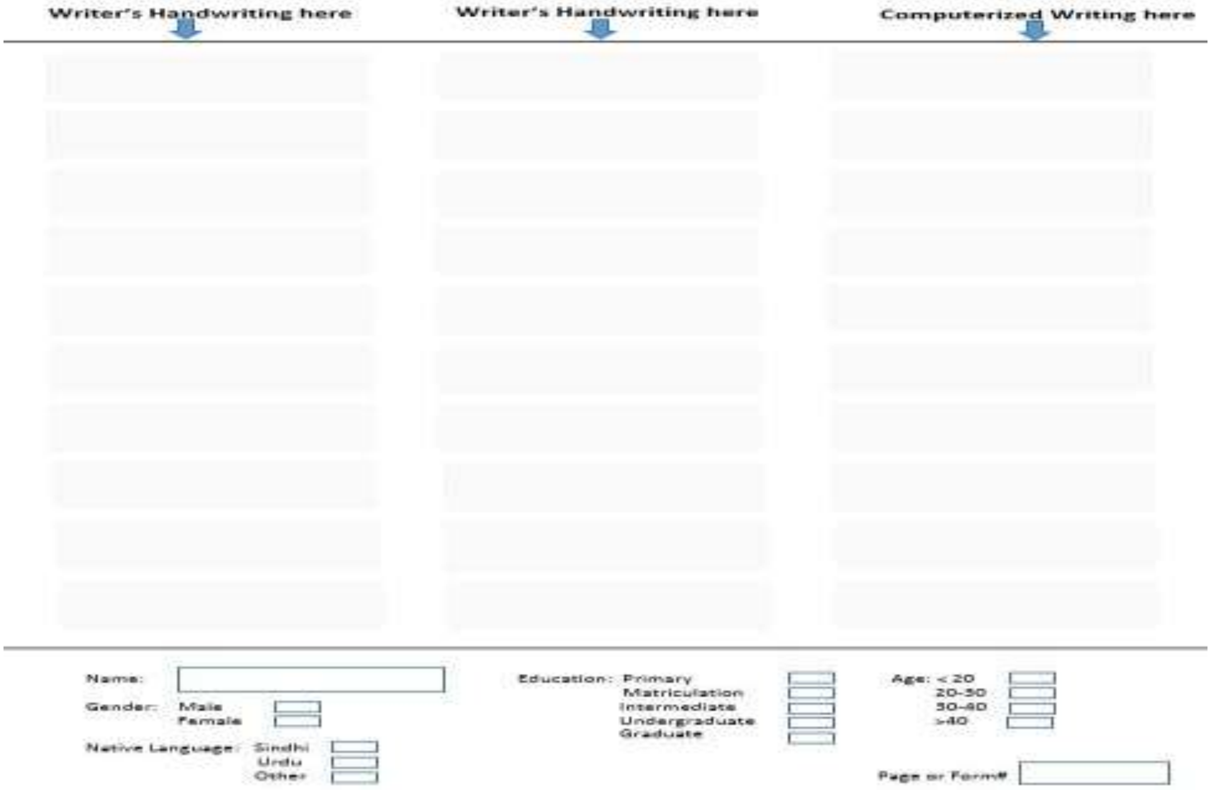
As there is no any Sindhi handwritten text recognition system is available, so there is no any database is available for the testing and training purpose. In order to construct a Sindhi image database for the testing and training of Sindhi handwritten text recognition, it is very necessary to collect the data, design the forms and in last which is necessary to invoke the techniques which can detect the number of words.

Data Collection:

For the creation of any database it is very necessary to collect the huge amount of data for the testing and training of any reliable recognition system. So we collect the data from the different writers with different writing styles. As some data are confidential, so we decided to collect the data in forms of names (boys, girls, cities, district, countries and capitals) a testing and training purpose. We prefer to choose the real world data rather than the artificial data because that would be easy for the writers to write. We have collected the different writing style with different pens from the different writers.

Designing of Forms for the database:

Forms creation is very necessary part of the database through which we collect the handwritten words for the training and testing of the database. Forms designing is the major portion of the database. Forms were designed using the MS Word and for the speedy work PHP Language has been used. There are two types of forms were designed the 1st from contains the three columns and 2nd form is very simple neat and clean page where user are asked to write in straight line. 1st form is totally technical which contain the three columns. Form is shown in figure 3. Form contains the three columns and ten rows and writer information at the bottom of the page. The first column contains the computerized or printed Sindhi writing, while the 2nd and 3rd column contains the user writing. The firs column may contains the computerized or printed cities names, districts names, human being names, capital and country name and in second and third columns user will ask to write the same names as they are in first column. The guidelines for the users writing is given on the top of the page.



Writer's Handwriting here | Writer's Handwriting here | Computerized Writing here

Name: □

Gender: Male □
Female □

Native Language: Sindhi □
Urdu □
Other □

Education: Primary □
Matriculation □
Intermediate □
Undergraduate □
Graduate □

Age: < 20 □
20-30 □
30-40 □
>40 □

Page or Form# □

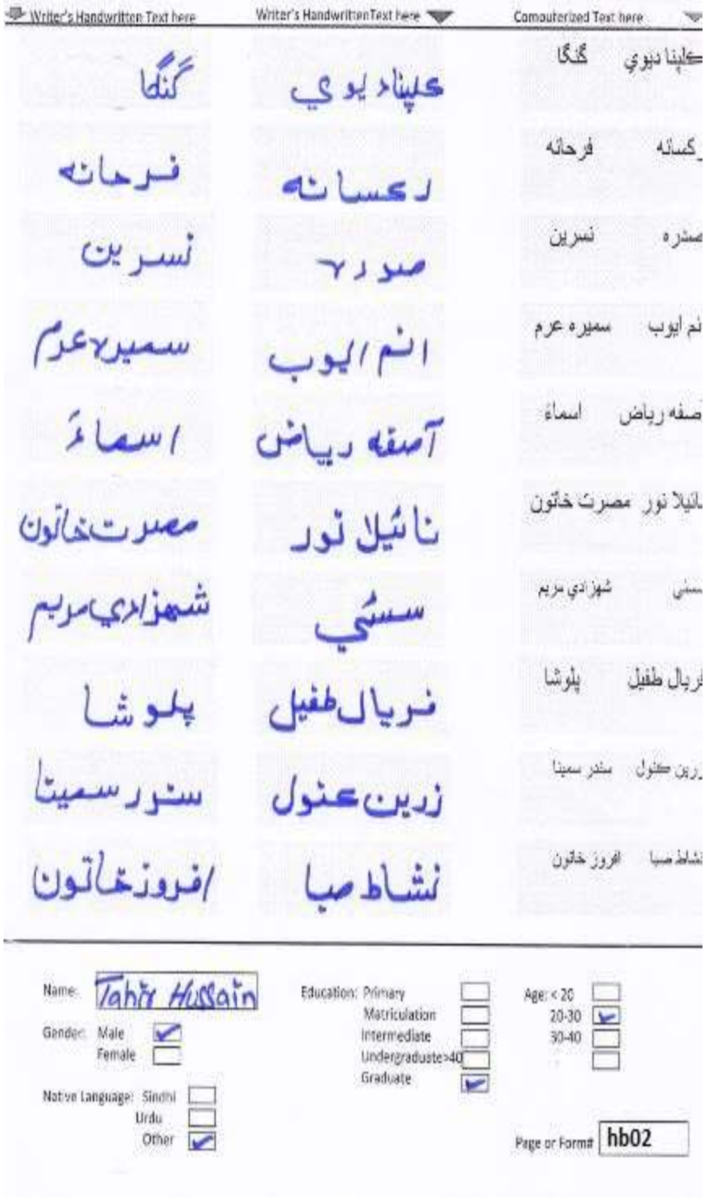
Here in the form we designed the column with little grey color rectangular boxes, where writers understand and sense the area of to writing. These rectangular boxes were designed using the MS word. A light grey shadow effects have been applied to make the rectangular boxes vivid and clear and border of rectangular boxes help the writers that where to write. After scanning process these rectangular boxes of grey color can be easily removed using the threshold values in MATLAB tool. The two straight lines have been inserted on the top of page and bottom of the page to check the skewness and direction of page during the scanning process. These rectangular boxes and straight lines on the forms are very necessary to perform the segmentation process on the data. The bottom block of the page contain the ground truth or additional information about the writers. Each writer were asked to write the five different forms of different names. At the bottom of the form the page number is mentioned, which is used to as a form identifier for the subsequent process.

Role of Pen for the creating of database:

Different pens were used for the writing on the forms. Writers were asked to write using the different colors of pen and try to make the image readable. The best result was given by dark black and blue pen. The nib of pen also matters a lot but thick writing gives better result. Ink distortion issues has been observed as well. It is recommended that connecting and joining the handwritten words should be given importance. It is also recommended that each word should be vivid and clearly written. The result of different pen writing is shown in figure 4(b).

Database Form Scanning:

Scanning of the form plays very important part of recognition rate. The handwritten form images should be scanned using the high dpi(dots per inch) as poor scanning compromises the recognition rate. The form images have been scanned at 100dpi, 150dpi and 300dpi. The images which were scanned at 300 dpi has very good results. Images were further cropped and finally stored for the creation and training of the database. Paper which is used for the collection of data is white. The page skewness and slope correction were corrected automatically using the two straight lines which has been marked at the top and bottom of the page. Form scanning is shown in figure 4(a).



Writer's Handwritten Text here	Writer's HandwrittenText here	Computerized Text here
گنگا	علپنا دیوي	ـکلپنا دیوي گنگا
فرحانه	رکسانه	ـکسانه فرحانه
نسرین	صورہ	ـصدره نسرین
سمیرہ عرم	انم /ایوب	ـنم ایوب سمیره عرم
/سماء	آصفه ریاض	ـصفه رباض اسماء
مصرت خاتون	نائیل نور	ـائیلا نور مصرت خاتون
شهزادي مریم	سسئي	ـسئي شهزادي مریم
پلوشا	فریال طفیل	ـریال طفیل پلوشا
سنور سمینا	زرین عنول	ـرین کنول سندر سمینا
/فروز خاتون	نشاط صبا	ـشاط صبا افروز خاتون

Name: Tahir Hussain

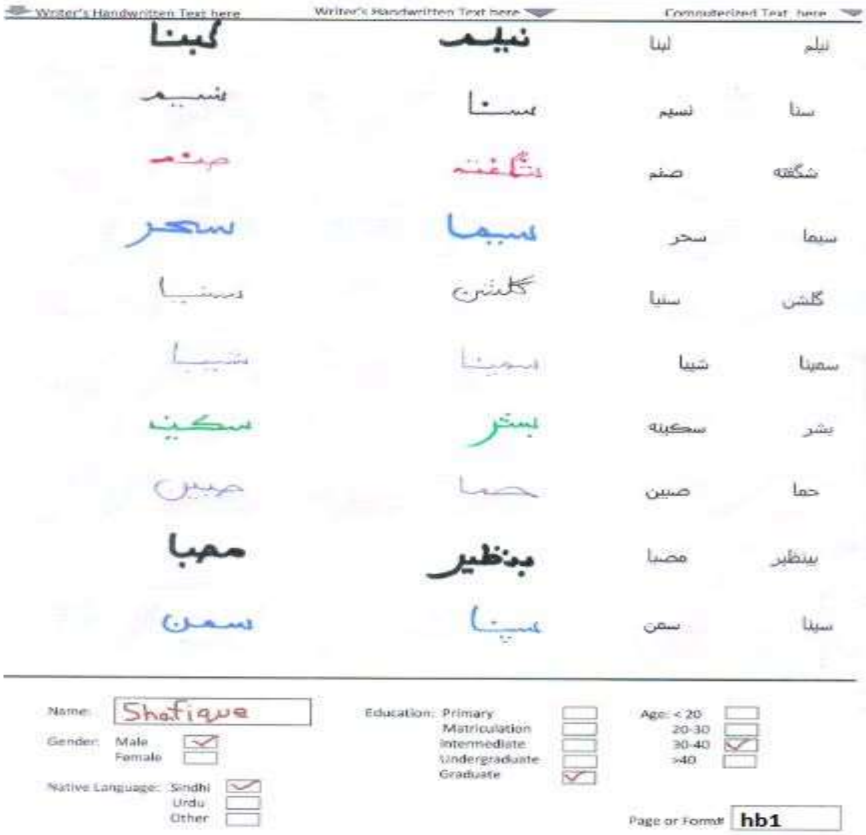
Gender: Male ☑
Female □

Native Language: Sindhi □
Urdu □
Other ☑

Education: Primary □
Matriculation □
Intermediate □
Undergraduate □
Graduate ☑

Age: < 20 □
20-30 ☑
30-40 □
>40 □

Page or Form# hb02



Writer's Handwritten Text here	Writer's Handwritten Text here	Computerized Text here	
لبنا	نیلم	لبنا	نیلم
شسیم	سنا	نسیم	سنا
صنم	شگفتہ	صنم	شگفتہ
سحر	سیما	سحر	سیما
دستیا	گلشن	ستیا	گلشن
شبیا	سمینا	شیبا	سمینا
سکینہ	بشر	سکینہ	بشر
صبین	حما	صبین	حما
مصبا	بینظیر	مصبا	بینظیر
سمن	سپنا	سمن	سینا

Name: Shafique

Gender: Male ☑ Female □

Native Language: Sindhi ☑ Urdu □ Other □

Education: Primary □ Matriculation □ Intermediate □ Undergraduate □ Graduate ☑

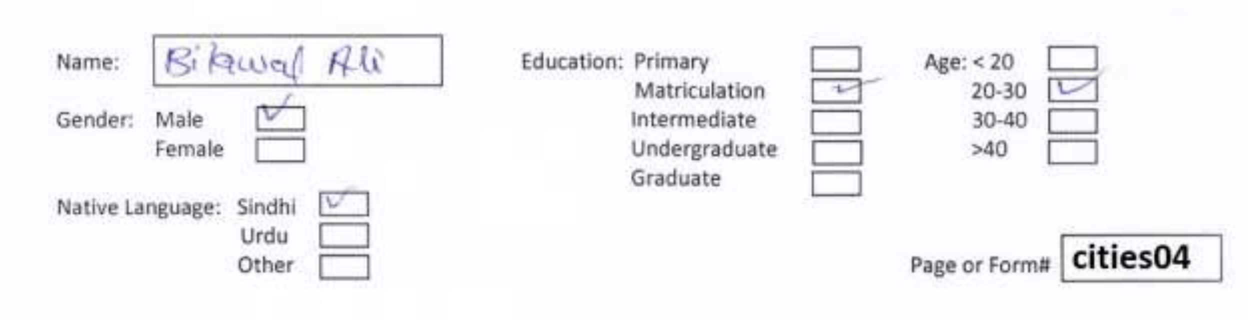
Age: < 20 □ 20-30 □ 30-40 ☑ >40 □

Page or Form# hb1

All the forms were stored in Bitmap(BMP) format. BMP files may be easily created from the existing pixels because it stored the pixels in array of memory. BMP format has higher acceptance and use no any compression technique. Bitmap files may be translated dot format output devices such as CRTs and printers. In BMP format pixels may be modified individually.

Writer's information:

Writer's information are written at the bottom of the page. Which is very necessary to maintain the statistics and record of data sets. Writer's information contains name, gender, education, age and language. This writer's information is necessary to construct writer sets from the collected data. We have collected the five different forms from each writer. Writers are categorized according the native language because writing styles varies according to the native language.



Name: Bikwal Ali

Gender: Male ☑ Female □

Native Language: Sindhi ☑ Urdu □ Other □

Education: Primary □ Matriculation ☑ Intermediate □ Undergraduate □ Graduate □

Age: < 20 □ 20-30 ☑ 30-40 □ >40 □

Page or Form# cities04

Structure for the Extraction of Data from the Image for database creation

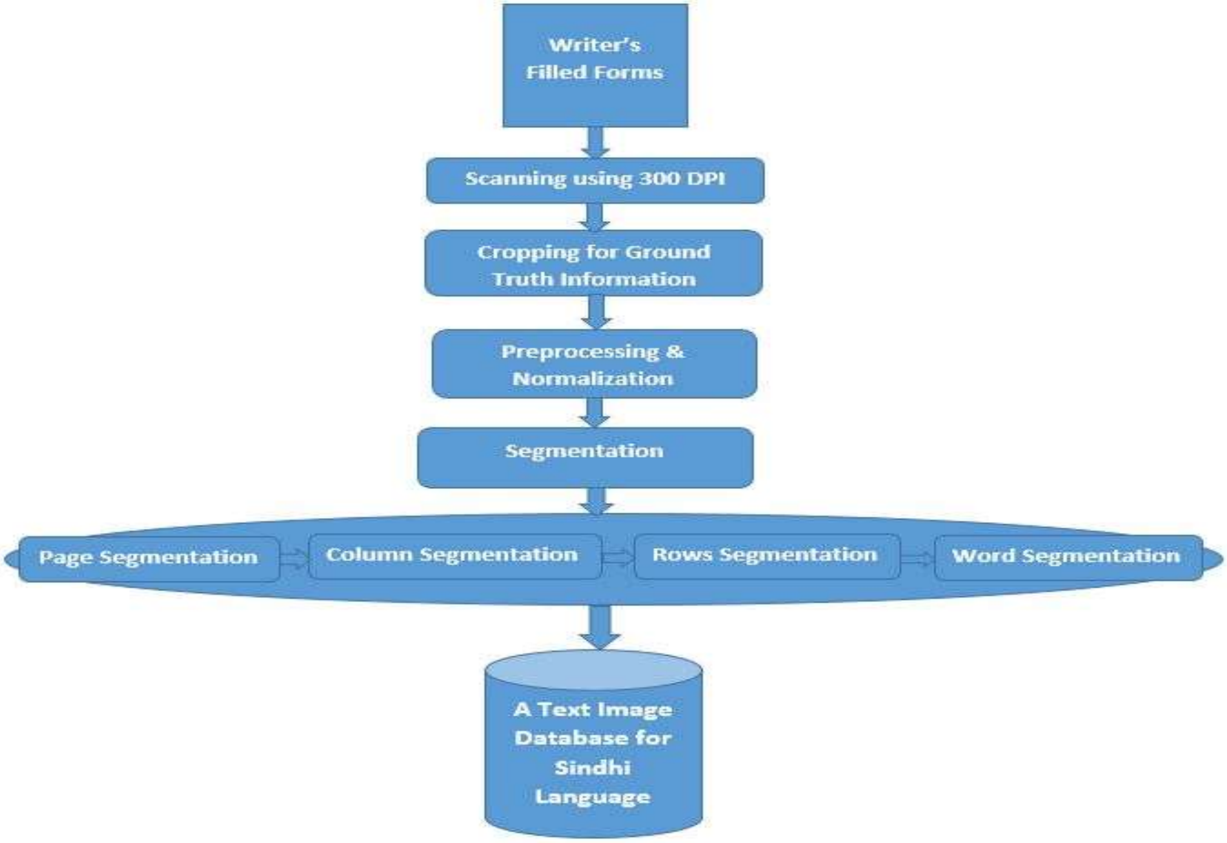


Figure 11: The Structure for the Sindhi Text Image Database

This is the basic structure of Sindhi Text Database. This is the basic techniques which are the basic model

Enhanced Segmentation Techniques for Sindhi Segmentation

A structure for the creation of Sindhi Text image is given in figure 11. The description of the stages are given bellow:

1- Preprocessing:

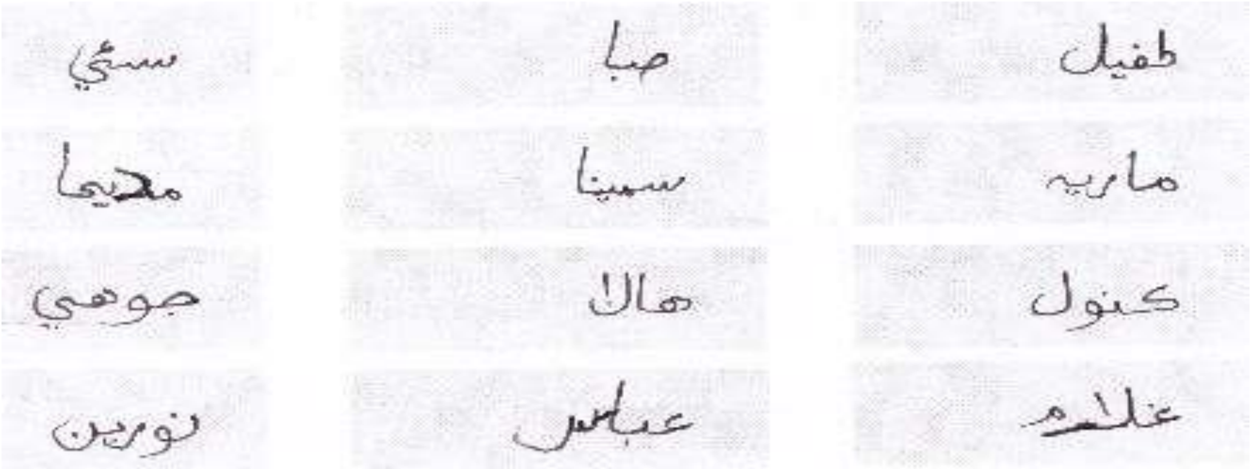
In the preprocessing the text is normalized. The stroke of pen is corrected. The main purpose software is normalization, which helps to remove variation of images and correct the text. The extra noise or dots have been removed by the dots.

2- Slope and Skew Correction:

In the slope correction the angle of base line is estimated and corrected using the horizontal projection and vertical histogram. The slope is calculated by finding the lowest and highest pixel in vertical scan line, it calculate the chain of pixels. Further the slope is corrected by putting the lines top and bottom of the page. Here the angle of the text is corrected using the horizontal projection and vertical histogram. In the slope correction estimation of the base line is performed by counting the position of black pixel of image and then vertical histogram is applied.

3- Grey Scale Effects

Grayscale effect was applied on the images as the text can be recognized and remaining noise can be removed easily. The Grey Scale Effects have been applied on the images in which the each pixel values is sampled. The basic function of grey scale effect find the intensity of information of each pixel. The each color pixel is mapped into black and white. Figure 6 showing the grey scale effects applied by MATLAB tool.



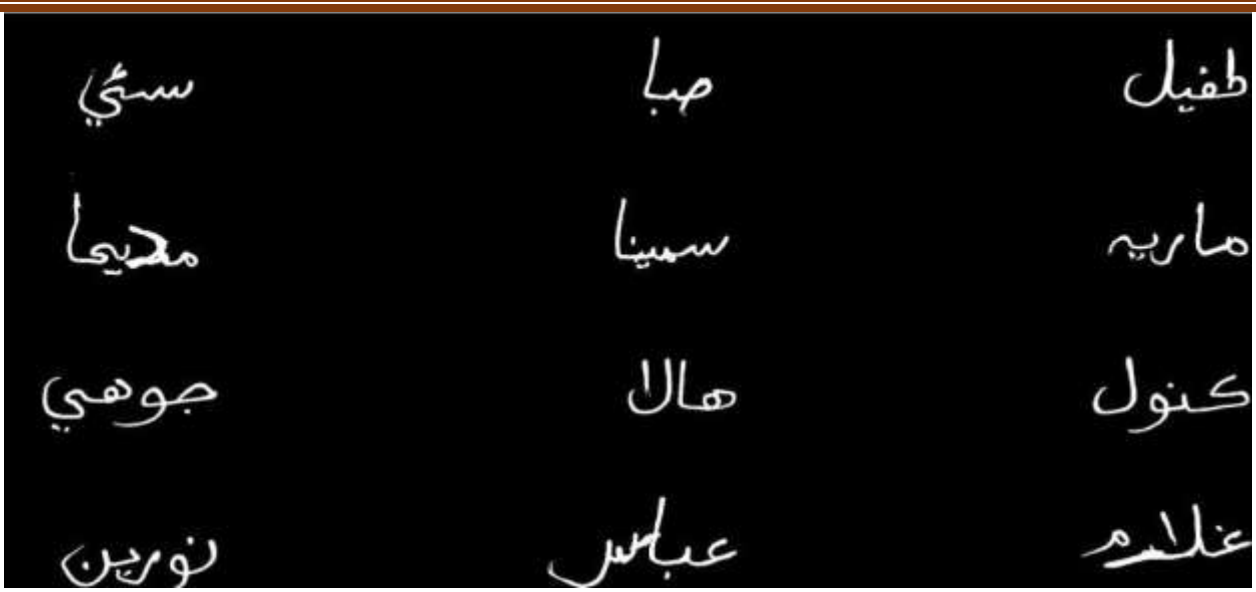
4- Binarization of Image

In the Binarization process the grey scale image is convert to black and white. The basic purpose of binarization is to extract the text from the image. Binarization helps to increase the quality of database. The Binarization is very important to increase the recognition rate of text. Since the quality of the binarization method employed to obtain the binary result depends on the type of the input image (scanned document, scene text image, historical degraded document etc. The extra noise and unnecessary dots from the images are easily removed by the executing the function of Binarization in MATLAB. The effects of binarization is shown in figure 7

طفیل
صبا
سسي
ماریہ
سمینا
مدیحا
کنول
هالا
جوهي
غلام
عباس
نورين

5- Final Segmentation for creation of database

Segmentations were used to separate the information from the form. Whole page was segmented using the 90 angle. Each line, column and finally words were segmented. Line segmentation has been performed by counting the free space between the lines and words. Different types segmentation is shown in figure 8.


طفيل
صبا
سمي
ماريه
سمينا
مديحا
كنول
هالا
جوهي
غلام
عباس
نورين

صبا
سمينا
هالا
عباس

كنول
هال
جوهي

كنول
هال
جوهي

Database Ground Truth information:

References:

Mahmoud, S. A., Ahmad, I., Alshayeb, M., & Al-Khatib, W. G. (2011). A database for offline Arabic handwritten text recognition. In *Image Analysis and Recognition* (pp. 397-406). Springer Berlin Heidelberg.

Al-Ma'adeed, S., Elliman, D., & Higgins, C. A. (2002). A data base for Arabic handwritten text recognition research. In *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on* (pp. 485-489). IEEE.

Alamri, H., Sadri, J., Suen, C. Y., & Nobile, N. (2008). A novel comprehensive database for Arabic off-line handwriting recognition. In*Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR* (Vol. 8, pp. 664-669).

Pechwitz, M., El Abed, H., & Märgner, V. (2012). Handwritten Arabic word recognition using the IFN/ENIT-database. In *Guide to OCR for Arabic Scripts*(pp. 169-213). Springer London.

Hull J. J., (1994). A database for handwritten text recognition research. Pattern Analysis and Machine Intelligence, vol. 16, no.5, pp 550-5544. IEEE

Al-Ohali, Y., Cheriet, M., & Suen, C. (2003). Databases for recognition of handwritten Arabic cheques. *Pattern Recognition*, *36*(1), 111-121.

El-Sherif, E. A., & Abdelazeem, S. (2007, July). A Two-Stage System for Arabic Handwritten Digit Recognition Tested on a New Large Database. In*Artificial Intelligence and Pattern Recognition* (pp. 237-242).

Mahmoud, S. (2008). Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models. *Signal Processing*,*88*(4), 844-857.

Wahab, M.; Amin, H. & Ahmed, F. (2009) : "Shape analysis of Pashto script and creation of image database for OCR". In:: . : Emerging Technologies, 2009. ICET 2009. International Conference on.S. 287-290

Jabril Ramdan, Khairuddin Omar, Mohammad Faidzul, Ali Mady. (2013): " Arabic Handwriting Database for Text Recognition": ScienceDirect Elsevier Procedia Technology 11 580-584

AbdelRaouf, A., Higgins, C. and Khalil, M. (2008). A database for Arabic printed character recognition, in A. Campilho and M. Kamel (eds), Image Analysis and Recognition, Vol. 5112 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 567–578.

Luqman, H., Mahmoud, S. A. and Awaida, S. (2014). KAFD Arabic font database, Pattern Recognition 47(6): 2231 – 2240.

Alana, G. A. (1993). "Sindhi Sooratkhati, fourth edn", Sindhi Language Authority, Hyderabad, Sindh, Pakistan.

Al-Badr, B. and Mahmoud, S. A. (1995). Survey and bibliography of Arabic optical text recognition, Signal Processing 41(1): 49 – 77.
URL: http://www.sciencedirect.com/science/article/pii/016516849400090M