SindhUniv. Res. Jour. (Sci. Ser.) Vol. 51 (04) 595-600 (2019)
http://doi.org/10.26692/sujo/2019.12.95
SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Dataset of Urduud1k from Natural Scenes

U. ZAKI, D. N. HAKRO, M. MEMON, F. H. KHOSO, K. U. R. KHOUMBATI, M. A. ZAKI* M. HAMEED, G. NABI

Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan

Received 08th February 2018 and Revised 19th July 2019

Abstract: In latest years research has drawn attention on text analysis in natural scenes. Databases play a significant part in the efficiency assessment of the algorithm for text recognition. A data set of natural scene text images in six distinct languages have recently been released in an International Conference on Document Analysis and Recognition (ICDAR).This dataset is for multi-languages except Urdu. In the natural images of the Urdu scene, there is an absence of a conventional Urdu text database. This research therefore mainly aims to build a database for Urdu text in natural scenes. The dataset is very large because there are 10 distinct cameras with distinct resolution, distinct angles and distinct range requirements for each picture captured by distinct light zone. The dataset comprises of Urdu words, ligatures and characters in natural scenes. The dataset contains 16k images of words, 32k ligatures and characters images. This dataset contains 1kimagesincluding signboard, a name of the store, banners and so on. In addition, the Urdu dataset is contrasted with the current data set including ICRAR 2003, ARASTI, Chars 74k, etc. The dataset includes many images from the natural scene so it can be used in natural environments to identify Urdu text.

Keyword: Urdu in Natural Scene; Dataset of Urdu Text; Urdu OCR; Detection and Recognition of Urdu

1. INTRODUCTION

In the natural scene, the text is widespread including signs, traffic signs, store names, etc. These texts are the significant indications for understanding the content of the picture and offer helpful and significant scene data (Sharma and Prakash, 2012). Text extraction from the natural scene comprises of two primary steps identification and recognition of text. A phase is called Text Detection, which is used to identify the text area of an image. The next stage is to recognize the text identified in the dataset. Text detection and recognition on paper base (Scanned documents) have traditionally been conducted. Thus, traditional OCRs (Optical Character Recognition) correctly execute scanned files on the white background but do not recognize and detect text from natural scenes. Due to text changes in dimensions, color, resolution, design, blurring issues, orientation, alignment, shape, texture, background, text geometry, lighting issue, contrasts with background texture and its identification from scenes of natural environments is challenging (Zaki *et al.* 2019).

Language is the principal sign of any nation's identity. The world has so many speak and written languages. Most images of natural scenes contain texts in several languages. It demonstrates that it is helpful to international visitors to translate and comprehend street signs, product labels and shop names when the picture of the naturally occurring scene is used for text recognition. Identification and recognition of text from natural scenes in many scripts, including Chinese, English as well as Arabic, but not done yet in Urdu because of a lack of databases.

Recognition of natural scene in latest years has become more attractive. Some study was carried out on individual characters and on the development of databases for text recognition from natural scenes. However, for natural scene recognition, Urdu does not have a dataset. The primary purpose of this study therefore is to produce the Urdu text dataset for the identification of texts in the natural scene and to enable other researchers to compare and propose various methods.

Urdu text detection is a difficult job due to the sensitive nature of the context which means that most Urdu characters have up to four different shapes according to the position isolated, center, initial and final, as illustrated in **(Fig. 1).**

Fig. 1: Position of Urdu Character "Meem"

*Department of Computer System Engineering, Mehran University of Engineering and Technology, Jamshoro,

Urdu consists of 38 basic characters (Krizhevsky, *et al..*, 2014). It does not have upper and lower case of characters but has non-joiner characters (can't combine with other characters) and joiner (can combine with others characters formed new word) as shown in **(Fig. 2)**. Ligatures indicate the connected character. Urdu word formed by the combination of minimum two and maximum four ligatures (Zaki, *et al..*, 2019). Some samples of Urdu word according to number of ligatures shown in **(Fig. 3)**.

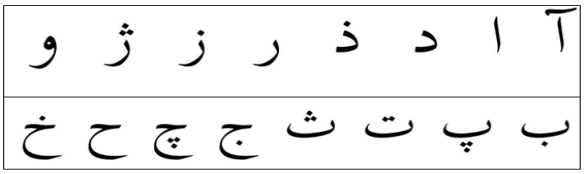

Fig.2: Example of Urdu Characters Non-Joiner &Joiner

Fig. 3: Example of Words According To Ligatures

The paper is organized in several sections. Section II, Presents the rationale for database collection. The existing dataset will be discussed in Section III. Section IV showed the characteristics of the proposed database and Section V showed the conclusion and future work.

The Rationale for the Collection of this Database

In the natural scene it is a very difficult task to accurately detect Urdu text due to the variation of texts in size, color, broken ligature, overlap of words, fluttering problem, texture, background, text geometry, lighting issue and background contrast as described in **(Fig.4).**

Fig. 4: Challenges of Urdu Text Detection

This research offers a new dataset of Urdu text images, particularly the signboard text with the city's name, the name of the store, and banners with different, distinct resolution mobile cameras. There is no dataset available for Urdu Text, according to the literature review. So, it's a novel database for Urdu script in natural scenes. This database targets the city name, cropped word and characters in natural scene images.

2. REVIEW OF RELATED WORK

Urdu is a Pakistani national language and has also been spoken fluently in various Indian states. Recent study has drawn interest in the evaluation and appreciation of Urdu papers (hand-written or printed). Contributions towards the development of Handwritten and Artificial dataset of Urdu script have already been made. (Malik and Khan (2005) proposed a system for online Urdu handwritten. 39 individual characters and 10 numbers were tested. (Hussain *et al.*, 2007) reposed online Urdu handwriting recognition. (Javed and Hussain (2009) developed Urdu OCR in nastalique script.

Lehal and Rana (2013) presents a system that recognized Urdu 9262 ligatures in nastalique writing style. (Naz *et al.* 2016) roposed a dataset consist of 10,000 text line, and 771,339 characters. (Sahabbbir and Siddiqi 2016) use specimens of 30 document images in the invariant size databases from the Center of Language Engineering (CLE).

(Israr-uddin *et al.* 2016) the hybrid offline Urdu documents method suggested. In these 306 texts, the scheme has been effectively segmented for 30 papers composed of 310 text lines. In these 6,811 ligatures, these lines are 7364 ligatures. The dataset for Urdu script in natural scenes is compared with the existing and standard dataset of natural scenes.

Different versions of the datasets are created and published for various languages in natural scenes including Arabic, English, Chinese, and much more. Below are some of the examples from this dataset as shown in Figure 5.Table 1 shows the captured images, trimmed word and character statistics.

a. Dataset for English

ICDAR 2003: In the Robust Readings Competition suggested by Lucas *et al.* (2005), the character dataset is frequently used. The dataset consists of 6100 training characters and 5400 test characters. It contains 509 images of texts in the scene, 999 words and 11615 characters.

Chars74K: (Campos 2009) presents database consists of 12503 total images for English character. These images are manually segmented on which for experimental purpose 7705 images are useful. The dataset has 64 classes containing numbers, upper case and lower case English alphabets. The datasets also includes 62992 synthetic and 3410 hand drawn English characters.

SVT: (Wang *et al.* 2011) presents a dataset for Street View Text consisting of 52 classes of 3796 character samples. It contains 350 images with 647 words and 3796 letters. This dataset is more difficult because it has variety of low resolution images having different font's style images, and also has low lighting images.

MSRA TD 500 (Yao *et al.* 2012) introduce an indoor and outdoor natural scene dataset captured by handy camera. The dataset comprises of 500 images, 300 of them for training and 200 for testing.

IIIT 5K-word Mishra *et al.* 2016) introduce IIIT 5K-word dataset. This dataset is divided into two sets. One set for training, having 9678 samples and other for testing having 15,269 samples for character. The database consists of 5000 images on which 3000 images used for testing and 2000 images are used for training. It is difficult due to font's variable size, color configuration and noise presence, fog, distortion, and variable lighting.

CIFAR-10: (Krizhevsky *et al.* 2014) introduces Canadian Institute for Advanced Research (CIFAR-10) dataset for text recognition in machine learning. It includes 10 classes, 6000 images per class and 60,000 images for color samples.

ICDAR 2015: (Karatzas *et al.* 2015) will present a new dataset of 1,670 images. It isa further dataset of current ICDAR 2003, ICDAR 2005 and ICDAR 2011**(**Karatzas *et al.* 2011) .

b. Dataset for Chinese

GB2312: (Yu and Wan 2016) suggested a scheme that uses Smart Phone to detect Chinese text from natural scenes. GB2312 is a 6763 normal Chinese character set used commonly between 3755. The system establishes seven popular font sizes for 3755 characters and eight distinct sizes. There are 56 samples per personality to be tested.

Chi-Photo & Pan-Chinese-Character: (Tian *et al.* 2016) planned two datasets for Chinese characters termed' Pan-Chinese-Character' and' Chi-Photo.' 248 images of natural scenes used for training and 239 for testing purposes by producing Chinese character pictures 3419 for training and 2763 for testing in Pan Chinese character. Instead, 343 natural scene pictures are used as Chi-Photo datasets in total. The dataset composed of 1115 Chinese character classes for 4476 character samples.

c. Dataset for Hindi Script

Chars74K: The ancient Dravidian language is Kannada. It has 49 fundamental characters, but can be coupled with more than 600 different classes vowels and consonants (Tounsi, *et al.* 2017). presents Chars74 K dataset which includes 4194 Kannada text images. These images are divided manually and 3345 images are helpful for experimental purposes.

DSIW-3K: (Murthy *et al.* 2018) presents the datasets of Devanagari Script collected form Signboards, advertisements on roads, road signs, shopping areas, and public places, such as park, etc. In the Devanagari script, the data collected not only on the printed device but also have handwritten text. The dataset consists of 23040 samples of machine printed characters and 30,355 samples the handwritten Devanagari characters.

IIIT-ILST dataset: (Murthy *et al.* 2018) suggested a dataset containing over 100 pictures of a natural scene for Telugu, Malayalam, Devanagari the three Indian writings. The dataset acquired by the camera from billboards, local markets, and banners, signals etc. Every script obtained from these pictures contains about 1000 words of text.

ISI-Bengali-Character: A dataset called ' ISI-Bengali-Character ' was suggested for the first time for Bengali. It is the sixth most common script in the globe as a language. There are 50 fundamental characters in Bengali in which 11 are vowels and the rest are consonants. This enhanced dataset have 40 samples per character class. It comprises of 260 images and 4280 manually designed samples, including 15 250 Bengali character samples.

d. Dataset for Arabic

ARASTI: A data set named ARASTI was suggested for the natural scene Arabic text recorded (Tounsi, *et al.* 2017). The Arabic script has 28 fundamental and cursive characters in nature. Some characters also have distinct forms (original, middle, final and isolated) depending on place. This is the first Arabic scene images dataset. The data set consists of 1687 images with text, 1280 images with Arabic words and 2093 images with Arabic character.

Artificial Arabic Text: (Zayene *et al.* 2015) suggested dataset on videos for Arabic artificial text. The dataset was gathered from 4 news chains in Arabic. It consists of 80 videos with 850,000 frames.

e. Other Datasets:

A translation application suggested for the use of a mobile camera to translate English into Spanish. 21,000 English word datasets were created by the scheme. The translation system requires 2.51 seconds.

An application that's translates Japanese script into English. In 119 natural images, 141 texts were tested (Watanabe, *et al.*, 2012).The application suggested that recognizes text from the natural scene and then translates into the required script. The system produced its own 400 pictures from Nokia Smartphone N900. A total of 100 pictures in which 58 words were discovered are used to test. 26 read properly and 20 properly translated (Petter, *et al.*, 217) For Artificial Urdu on Videos, (Raza and Siddiqui, 2012) proposed a novel Data set of Artificial Urdu on video images. This database contains 1000 video images collected from 19 channels selected as sports, entertainment, news, religious, and business channels. The dataset contains 3339 numbers and about 23, 833 Urdu words in the images collected from videos.

(Mirza, 2018) work on Urdu artificial texts appearing in the video of hundreds of TV channels as shown in **(Fig.5-6).** Develop a dataset from various news channels consisting of 2000 video frames.

Fig. 5: Sample of datasets images in different languages: (a) ICDAR 03 (b) Chars74K (Eng) (c) MSRA-TD 500 (d) Chars74K (Kannada) (e) IIIT-ILST (f) DSIW-3K (g) ISI-Bengali-Character (h) ARASTI (i) Artificial Arabic Text

Fig. 6: images collected from videos (Mirza, 2018)

Table 1: Comparisons & Summary of Datasets of Natural Scene

Dataset Language	Dataset	Captured Images	Segmented Word	Segmented Characters
English	ICDAR 2k3	509	999	11615
	Chars74K	7705	----	12503
	SVT	350	647	3796
	MSRATD 500	500	----	----
	IIIT5K-word	5000	----	----
	CIFAR-10	60,000	----	----
	ICDAR 2k15	1,670	----	----
Chinese	GB2312	----	----	6763
	Chi-Photo	343	----	4476
	Pan-Chinese-Character	487	----	6182
Hindi Script	Chars74K Kannada	4194	----	----
	DSIW-3K Devanagari	----	----	53395
	IIIT-ILST Telugu, Malayalam & Devanagari	100	1000	----
	ISI-Bengali-Character	260	----	15 250
Arabic	ARASTI	1687	1280	2093
	Artificial Arabic	850,000	----	----
Japanese		119	141	----
Proposed Dataset	**UD1K**	**1000**	**16,000**	**32,000**

Proposed Dataset for Urdu Text in Natural Scene

The objective of this study is to produce an Urdu text dataset for the recognition of text in the natural scene and provide researchers with the possibility of comparing distinct methods and suggested them. It includes 1,000 pictures of natural scene from town names, store names, banners, announcements, and so on, from Pakistan (Province of Sindh).The images will be captured in various zones of light, separate cameras with various resolution, angles and distances. More than 10 samples per picture are therefore available. **(Fig. 7)** shows some samples of the suggested dataset.

Fig. 7: Proposed Dataset Samples (a) City Name Signboard (b) Shop Name Board

A customized approach was created to generate a giant database of Urdu from natural landscape using MatLab R2015a. The system load pictures of scenes with different types of colors and fonts angle and generate text images in the binary format. Furthermore, the system saves the Urdu words, ligature and characters in separate folder. **(Fig. 8(a)** shows a snapshot of customized application and **(Fig. 8(b)** shows the flow diagram of the approach.

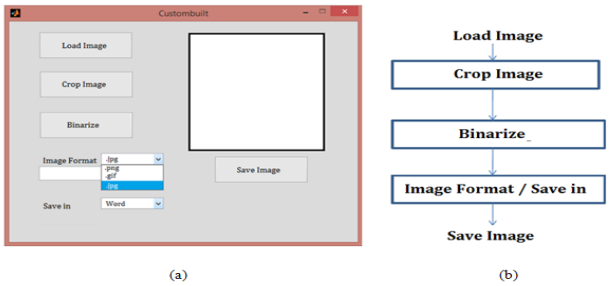

Fig. 8: Customized Applications: (a) Snapshot (b) Flow Diagram

Urdu words Image Dataset: The dataset comprises 16,000 Urdu cropped word images in natural scenes. This dataset is useful in Urdu entity recognition systems and also help the researcher to evaluate and develop word detection algorithms from natural scenes. Some samples of Dataset in Urdu are given in **(Fig. 8).**

Fig. 9: Samples of Urdu word cropped from Natural scene Images

Urdu Ligature and Characters Segmented Image Dataset: Urdu script has diversified word dataset because many words are formed by the combination of Ligatures. Hence, we enhanced our dataset by adding ligatures segmented images as shown in **(Fig. 9).**

Fig. 10: Ligature Segmented Images Sample

Urdu characters in images are segmented manually according to their distinct forms. Urdu is made up of 38 fundamental characters, both joiner and non-joiner. Hence, 40 classes of these characters are created according to shape (Isolated, Initial, Middle and Final) shown in **(Fig 10).** Hence we created a dataset contain 32,000 Urdu Ligature and Character Images.

Fig. 11: Character Segmentation Images Sample

Recently, (Chandio *et al.* 2018) proposed a Urdu Character dataset of 18000 cropped Urdu characters only from natural scenes but it is only for isolated characters. The proposed dataset not only consists of Urdu cropped words but also have a wide range of ligatures and character segmented images in natural scenes. So the proposed system compared with other Urdu character recognition system in **(Table 2).**

Table 2: Comparison of proposed system with existing Urdu Signboard System

Methods	Dataset
Chandio et al. [29]	18,000 cropped Character of Urdu
Proposed System	Huge dataset (16,000 words and 32,000 segmented ligature and characters)

3. CONCLUSION

A dataset of Urdu from nature is presented in this document. This is a perfect natural image dataset for Urdu Text. The dataset comprises of the separate images in natural scenes of the Urdu words, ligatures and characters. The dataset is limited to 1,000 natural landscape images gathered from the names of the shop, banners, etc. Furthermore, the Urdu dataset is likened to current available dataset including ICRAR 2003, ARASTI, Chars 74k, and many others. In future, the dataset will further be increase by adding more words and character classes. Furthermore, there will be a dataset for Pakistan that may include all the other languages datasets that fluently spoken in Pakistan.

REFERENCES:

Chandio, A. A., M. Pickering, and K. Shafi, (2018) "Character classification and recognition for Urdu texts in natural scene images," *Int. Conf. Comput. Math. Eng. Technol. Inven. Innov. Integr. Socioecon. Dev. iCoMET 2018 - Proc.*, vol. 2018-Janua,1–6.

De Campos, T. E., and M. Varma, (2009) "Character Recognition in Natural Images," *Visapp (2)*, 273–280.

Husain, S. A., A. Sajjad, and F. Anwar, (2007) "Online Urdu Character Recognition System," *MVA2007 IAPR Conf. Mach. Vis. Appl.*, 1-7.

Israr S. K., U. Din, and Zumra Malik, (2016) "Line and Ligature Segmentation in Printed Urdu Document Images Line and Ligature Segmentation in Printed Urdu Document Images," *J. Appl. Environ. Biol. Sci.*, vol. 6, no. March, 114–120.

Javed S. T. and S. Hussain, (2009) "Improving Nastalique-Specific Pre-Recognition Process for Urdu OCR,"

Kai Wang, B. Babenko, and S. Belongie, (2011) "End-to-end scene text recognition," *2011 Int. Conf. Comput. Vis.*, 1457–1464.

Karatzas D. *et al.*, (2015) "ICDAR competition on Robust Reading," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2015-Novem, 1156–1160.

Lucas S. M. (2005) "ICDAR 2003 robust reading competitions: entries, results and future directions," 105–122.

Lehal G. S. and A. Rana, (2013) "Recognition of Nastalique Urdu ligatures," 1Pp.

Malik S.and S. A. Khan, (2005) "Urdu online handwriting recognition," *Emerg. Technol. 2005. Proc. IEEE Symp.*, 27–31.

Mishra, A., K. Alahari, and C. V Jawahar, (2016) "Enhancing energy minimization framework for scene text recognition with top-down cues ☆," *Comput. Vis. Image Underst.*, vol. 145, 30–42.

Mathew, M., M. Jain, and C. V. Jawahar, (2018) "Benchmarking Scene Text Recognition in Devanagari, Telugu and Malayalam," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 7, 42–46.

Mirza, A. (2018) "Urdu Caption Text Detection using Textural Features," 70–75.

Naz, S., S. Bin Ahmed, R. Ahmad, and M. I. Razzak, (2016) "Zoning features and 2DLSTM for Urdu text-line recognition," *Procedia Comput. Sci.*, vol. 96, 16–22.

Petter, M., V. Fragoso, M. Turk, and C. Baur, (207) "Automatic text detection for mobile augmented reality translation.

Raza A. and I. Siddiqui, (2012) "A Database of Artificial Urdu Text in Video Images with Semi-Automatic Text Line Labeling Scheme," *4th Int. Conf. Adv. Multimed.*, 75–81.

Sharma S. and J. Prakash, (2012) "A Survey of Image to Text Detection Methodology," *Int. J. Adv. Res* vol. 3, no. 1, 46-49.

Shabbir S. and I. Siddiqi, (2010) "Optical Character Recognition System for Urdu Words in Nastaliq Font," *Inf. Emerg. Technol. (ICIET), 2010 Int. Conf.*, vol. 7Pp.

Shahab, A., F. Shafait, and A. Dengel, (2011) "ICDAR 2011 Robust Reading Competition Challenge 2 : Reading Text in Scene Images.

Tian S. (2016) "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognit.*, vol. 51, 125–134.

Tounsi, M., I. Moalla, and A. M. Alimi, (2017) "ARASTI : A Database for Arabic Scene Text Recognition," 140–144.

Watanabe, Y., K. Sono, K. Yokomizo, and Y. Okada, (2012) "Translation Camera On Mobile Phone Yasuhiko," 177–180.

Yao, C.. X. Bai, W. Liu, Y. Ma, and Z. Tu, (2012) "Detecting texts of arbitrary orientations in natural images," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 8, 1083–1090.

Yu B. and H. Wan, (2016) "Chinese Text Detection and Recognition in Natural Scene Using HOG and SVM," no. Itms, 1–5.

Yu, B. and H. Wan, (2016) "Chinese Text Detection and Recognition in Natural Scene Using HOG and SVM," *DEStech Trans. Comput. Sci. Eng.*, no. itms.

Zaki U. (2019), "Implementation Challenges in Information Retrieval System," *Sindh Univ. Res. Journal-SURJ (Science Ser.*, vol. 51, no. 2, 339-344.

Zayene, O., J. Hennebert, S. Masmoudi Touj, R. Ingold, and N. Essoukri Ben Amara, (2015) "A dataset for Arabic text detection, tracking and recognition in news videos- AcTiV," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, 996–1000.

Zaki, U., D. Hakro, K. R. Khoumbati, M. Zaki, and M. Hameed, (2019) "Issues & Challenges in Urdu OCR," *Univ. Sindh J. Inf. Commun. Technol.*, vol. 3, 1, 42–49.