Mehran University Research Journal of Engineering & Technology
Vol. 39, No. 1, 47-54, January 2020
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.2001.05

# A Database for Urdu Text Detection and Recognition in Natural Scene Images

ASGHAR ALI CHANDIO*, MEHWISH LEGHARI**, MUKHTIAR AHMED MEMON**, MEHJABEEN LEGHARI***, AKHTAR HUSSAIN JALBANI**

RECEIVED ON 20.08.2018, ACCEPTED ON 29.09.2018

## ABSTRACT

**This paper describes a novel database for Urdu Text detection and recognition in natural scene images. Many standard benchmarks for Latin text have been published, where remarkable classification and recognition techniques for text extraction in natural scenes are proposed. Recently, a dataset for multi-language text in natural scene images has been published by the International Conference on Document Analysis and Recognition (ICDAR). This dataset contains natural scene images in six different languages including Arabic, Korean and Chinese texts. Currently, there is no any dataset available for Urdu text in natural scene images. Therefore, the main objective of this paper is to create a novel dataset of Urdu text in natural scene images and provide to the research community to develop and evaluate state-of-the-art algorithms for text localization and recognition. The dataset consists of cropped words and segmented character images in natural scenes. All the characters are manually segmented from the captured images. All the images are captured in varying lighting conditions, low resolution, occlusions and perspective conditions. The dataset consists of 8000 cropped Urdu word-images and 16000 segmented Urdu character-images in different forms (isolated, initial, medial and final). The dataset is further increased by synthetically generating Urdu characters and putting on the real background images. The dataset is compared with the recently published Arabic natural scene datasets and Latin text datasets including ARASTI, ICDAR03 and Chars74k. The proposed dataset contains more natural scene images as well as segmented characters and cropped words, which show that the dataset can be used as a benchmark for recognizing Urdu text in natural scene images.**

**Key Words: Urdu Scene Text, Urdu Text Scene Charcter Recognition; Urdu Scene Dataset; Synthetic Urdu Scene Text**

## 1. INTRODUCTION

Text recognition in natural scene images has become a useful and challenging task in many real world applications. The text within natural scene images contains much valuable information, which is helpful to interpret the world and understand the other textual cues. It is one of the common ways of the commination. Text extraction in natural images is generally divided into two phases: detection and recognition. In

Authors E-Mail: (a.chandio@student.adfa.edu.au, mehwish.leghari@scholars.usindh.edu.pk, mukhtiar.a@gmail.com, legharimehjabeen@usindh.edu.pk, jalbaniakhtar@gmail.com)
* School of Engineering and Information Technology, University of New South Wales, Australia
** Department of Information Technology Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah Pakistan
*** Department of Information Technology, University of Sindh, Pakistan

detection, the image is checked if it contains text or not and in recognition, the detected text is converted into machine-readable form. Text recognition has traditionally been performed from scanned documents, where the text is usually in black-and-white, plain background and line based paper environment. In scanned documents, the text usually appears in consistent font type, size, color, style and fixed lines. Therefore, the Optical Character recognition (OCR) systems perform very well and accurate on these scanned documents. However, these OCR systems fail when applied to read text in natural scene images due to various challenges including background complexities, un-even lighting conditions, low resolution, blur, occlusion, variations in font size, type, color orientations and many more present in natural scene images. The natural scene images also contain many other objects whose structure resemble with the text, which make the recognition process further complex.

So far, most of the research works in scene text detection and recognition have been done for Latin text, where state-of-the-art results have been reported in several competitions i.e., Robust Reading Competition [1-6]. This is mainly due to the availability of many standard benchmarks including ICDAR [1-5], SVT [7], Chars74k [8], IIIT-5k [9], MSRA-TD500 [10] and many more. However, there are more than 100 languages commonly written and spoken around the world. Many natural scene images contain text in more than one language as well. This shows that if text recognition in natural scene images is carried for other languages, then it could be helpful for foreign tourists to translate and understand what is written on road signboards, shop names, advertisement banners and product labels.

Recently, some research work for isolated Arabic and Urdu character recognition in natural scene images has been reported and a dataset for Arabic scene text recognition has been developed [11]. This is the first benchmark for Arabic character recognition in natural images. A baseline research work has been done by [12][13] for isolated Urdu character recognition in natural scene images. However, no any dataset is available for Urdu text recognition in natural scene images. The availability of the standard datasets is important to evaluate existing state-of-the-art algorithms and to train and test machine learning classifiers for scene text recognition. Therefore, the main objective of this research is to capture natural images, develop a scene text dataset for Urdu text and make it available to the research community to propose and compare different techniques.

In Urdu scripts, the text is written from right to left direction. The text is written in different writing styles with or without dialectics. When present in natural scene images, the stroke width of each character is different at starting, medial and ending positions. Depending upon the position of the character within the word, each character can have more than one shape such as: initial, middle, final or isolated as shown in ((Fig 1. a), (Fig 1. b), (Fig 1. c) and (Fig 1. d).

In Urdu scripts, the words are formed by joining two or more characters in different shapes. Usually, in Urdu text, the words in natural scene images overlap with each other. Therefore, the automatic segmentation of these words into individual characters is sometimes impossible. Some of the source images and the manually segmented words are shown in Fig. 2 and Fig. 3.

In this paper, a novel dataset of Urdu images in natural scenes with cropped words and segmented characters containing Urdu text in advertisement banners, road sign

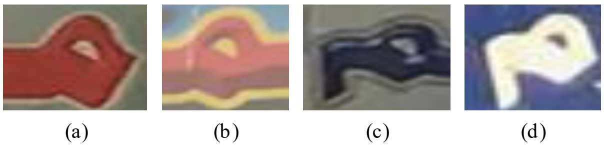


*FIG 1. INITIAL, MEDIAL, FINAL AND ISOLATED FORMS OF URDU CHARACTER 'NOON' WITHIN A WORD*

names, shop names, hoardings etc. captured with mobile camera is developed. To the best of our knowledge, this is the first dataset of Urdu text in natural scenes. While, some of the characters of Urdu text may occur very infrequently in natural scenes and other occur commonly, it is therefore investigated either to use synthetically generated characters with different fonts, sizes and styles or online hand-printed characters taken on the screen of smart devices to substitute the training data. To generate a synthetic dataset, more than eighty fonts of Urdu text are tried with different font sizes and styles (bold, italic and regular). The dataset is mainly targeted for isolated character and cropped word recognition in natural images. It will further be increased for whole image text detection and end-to-end text extraction algorithms.

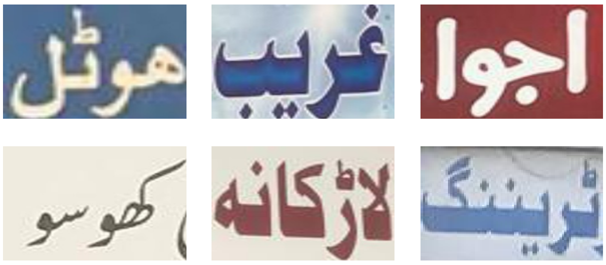


*FIG 2. SOME URDU TEXT WORD-IMAGES WHICH ARE FORMED BY JOINING TWO OR MORE CHARACTERS*

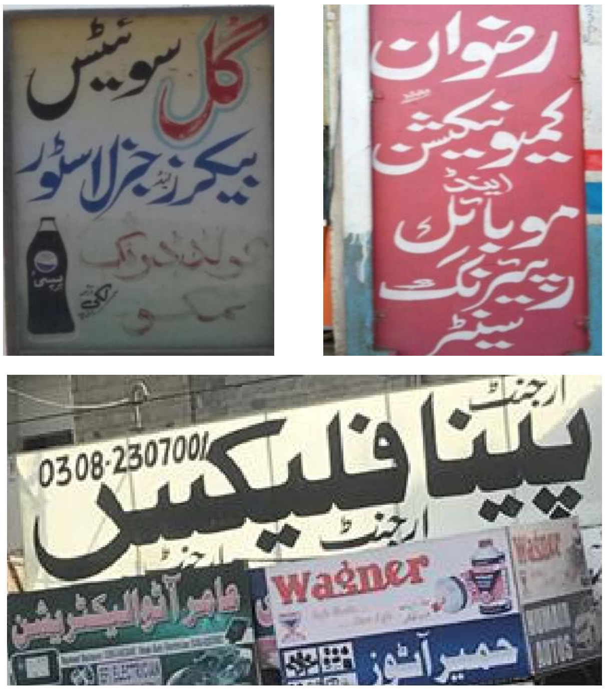


*FIG.3. SOME SAMPLES OF URDU TEXT IN NATURAL SCENE IMAGES WHERE THE TEXT IS OVERLAPPING WHICH IS DIFFICULT TO SEGMENT AUTOMATICALLY*

The rest of the paper is organized as follows: the next section describes the related existing datasets of other scripts. Section III highlights the proposed dataset, the segmented characters and words. Section IV explains the characteristics of the synthetic dataset and section V describes the concluding remarks and some possible future enhancements in the dataset.

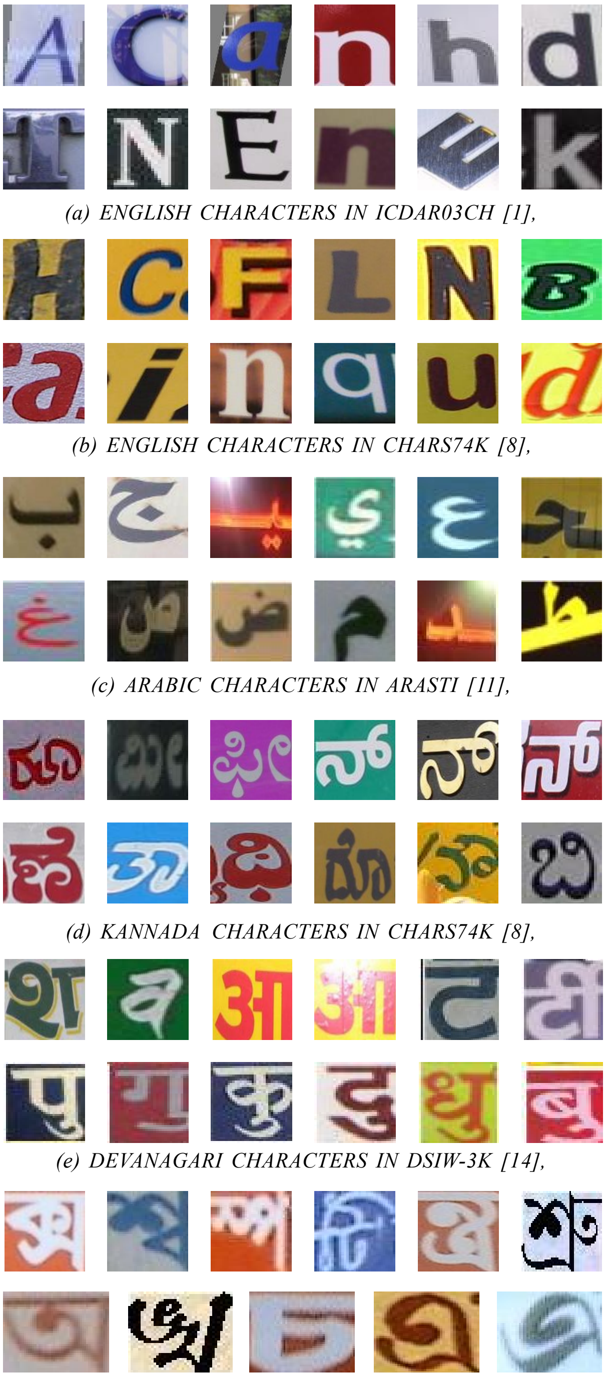
## 2. RELATED WORK

A number of standard benchmarks for scene text images are published by ICDAR. Several character image datasets in natural scene images for many languages including English, Arabic, Bengali, Kannada and Devanagari are generated and published. Some of the samples of the character images from these datasets are shown in Fig 4.

Fig 4. Samples of natural scene character images in different languages: (a) English characters in ICDAR03CH [1], (b) English characters in Chars74k [8], (c) Arabic characters in ARASTI [11], (d) Kannada characters in Chars74k [8], (e) Devanagari characters in DSIW-3K [14], (f) Bengali characters in [15].

ICDAR03CH dataset: it is one of the most commonly used natural scene English characters dataset in Robust Reading Competition. This dataset contains 509 natural scene images and 6185 character images for training and 5430 for testing. The character images are composed of 62 classes including digits 0-9, upper case and lower-case English letters.

Chars74K dataset: this dataset contains 1922 natural images with English and Kannada text [8]. A total of 12503 English character images are manually segmented out of them 7705 character images are used in experimentations and others are discarded due to low resolution, noise or occlusion. For Kannada text, a total of 4194 characters are manually segmented, out of them 3345 character images are used in experimentation and others are considered as bad images. This dataset also contains hand-printed characters for English and Kannada scripts as well as 62992 synthetic English characters.



(a) ENGLISH CHARACTERS IN ICDAR03CH [1],

(b) ENGLISH CHARACTERS IN CHARS74K [8],

(c) ARABIC CHARACTERS IN ARASTI [11],

(d) KANNADA CHARACTERS IN CHARS74K [8],

(e) DEVANAGARI CHARACTERS IN DSIW-3K [14],

K) BENGALI CHARACTERS IN [15].

FIG 4. SAMPLES OF NATURAL SCENE CHARACTER IMAGES IN DIFFERENT LANGUAGES

ARASTI dataset: this dataset contains the Arabic natural scene images and is called as ARAbic Scene Text Image (ARASTI) dataset [11]. A set of 1687 natural scene images is captured which is then segmented into 1280 Arabic scene word-images and 2039 character-images. The character-images are segmented into different shapes (isolated, initial, medial and final) within the words. This is the first dataset of the Arabic text in natural images. Another dataset of Arabic characters in natural scene images is proposed in [16], but this dataset contains very few numbers of images and the total segmented character images are also very low. Furthermore, the deep learning based data augmentation technique is used to enhance the size of the dataset by rotating each character-image at five different orientations.

Devanagari scene character dataset: this dataset contains the segmented character images of Devanagari text [14]. The dataset contains images of signboard names, advertisement banners, shop names, hoardings, etc.

Bengali scene character dataset: a Bengali scene character dataset proposed in [15] contains 15250 Bengali natural scene character images. The dataset also contain the Bengali numerals. A total of 260 natural scene images are captured from the road side signs and streets of west Bengal.

## 3. PROPOSED DATASET

The proposed dataset is compared with the currently available character datasets in natural images and the statistics of the number of images, cropped words and segmented characters is shown in Table 1. The dataset contains more natural scene images as well as cropped words and characters, which shows that this dataset can be used as a benchmark for the Urdu text in natural scene images. The existing datasets for Urdu text recognition are only limited to the printed scanned documents, handwritten and artificial text in still and video images. For artificial Urdu text, a dataset of 1000 video images in developed in [17]. All the images are captured from 19 different Urdu news and sports channels. A semi-automatic text line labeling method is also performed and the results are compared with the manual labeling procedure. The artificial text is not complex and challenging than the natural scene text, as this text has not big variations in text size, font type, style, color and alignment. Contrary to artificial text, the text in natural scene images has many variations in font size, type, color, and alignment as well as writing styles. The natural scenes usually have multiple objects in the background, which make the text more complex than artificial text. Therefore, the artificial Urdu text dataset proposed in [17] cannot be used for text recognition in natural scene images.

The major contribution of this paper is to provide a standard dataset to the research community for algorithm development and evaluation of the state-of-the-art techniques for Urdu text recognition in natural scene images. The dataset consists of 2200 natural scene images containing Urdu text, captured with mobile camera in different cities of Sindh province of the Pakistan. Most of the images captured are of signboards, shop sign names, advertisement banners, hoardings, etc. The images are captured in various lighting conditions, have different resolutions, and may contain blur and small text size. Some of the images of the proposed dataset are shown in Fig. 5.

Urdu Scene Character Dataset: Individual characters are manually segmented from the images in different shapes and a dataset of 16000 cropped characters is created. Each character image has a fixed width and height of 48x48 pixels and a total of 69 classes of 38 Urdu characters are obtained with all positions (isolated, initial, medial and final) within a word. Some of the samples of character dataset are shown in Fig. 6.

Each character class has unbalanced number of samples because some characters are not frequently used in text and some are more commonly used. Therefore, each character class has 30 to 1580 numbers of samples. To overcome the problem of unbalanced classes, a synthetic dataset of Urdu characters is created. The details of the synthetic dataset are described in the next section.

Cropped Urdu Word Image Dataset: a dataset of 8000 cropped word-images in natural scenes is also developed.

TABLE 1. SUMMARY AND COMPARISON OF DIFFERENT NATURAL SCENE TEXT IMAGES

| Dataset | Scene Images | Cropped Words | Segmented Characters |
|---|---|---|---|
| ICDAR2003 [1] | 509 | 999 | 11615 |
| Chars74K English and Kannada [8] | 1922 | ÑÑ | 12503 English and 4194 Kannada |
| Street View Text [7] | 350 | 904 | ÑÑ- |
| IIIT-5K [9] | ÑÑ | 5000 | ÑÑ- |
| Bengali Dataset [15] | 260 | ÑÑ | 15250 |
| Devanagri Dataset [14] | ÑÑ | ÑÑ | 3000 |
| Arabic Dataset [11] | 374 | 1687 | 2093 |
| English Arabic Scene Text [16] | Ñ- | ÑÑ | 540 |
| Urdu Dataset | 2200 | 8000 | 16000 |

This cropped word-image dataset can be used to develop and evaluate word-spotting algorithms. Some of the samples of Urdu word image dataset are shown in Fig. 7.

## 4. SYNTHETIC URDU CHARACTER DATASET

Synthetic data is an alternative way to be used for training the deep learning algorithms. Recently, very large synthetic datasets with synthetic text in natural scene images have been generated and the deep convolutional neural networks are trained in [18] and [19]. A synthetic dataset of 40250 Urdu character images is created, where the characters are placed at random positions with a background images. For background images, the datasets of [20-22] are downloaded and used. The font size is randomly selected from a range of values between 30 to 80 points and a font type is also selected from a list of 80 different Urdu fonts. Different properties of the fonts (italic, bold and regular) have been used. The color of the text is randomly generated from a list of RGB values between 0-255. In the proposed synthetic Urdu character dataset, the real background images are used, and the dataset resembles with the real natural scene datasets. Therefore, the deep learning algorithms can be trained on it, which can generalize to real natural scene images (non-synthetic). With synthetic engine, it is possible to generate a larger dataset along with labels automatically which can then be used to train the data-hungry deep neural networks. Some of the images of the synthetic Urdu characters are shown in Fig 8. This dataset will further be increased and extended to synthetic word-images.

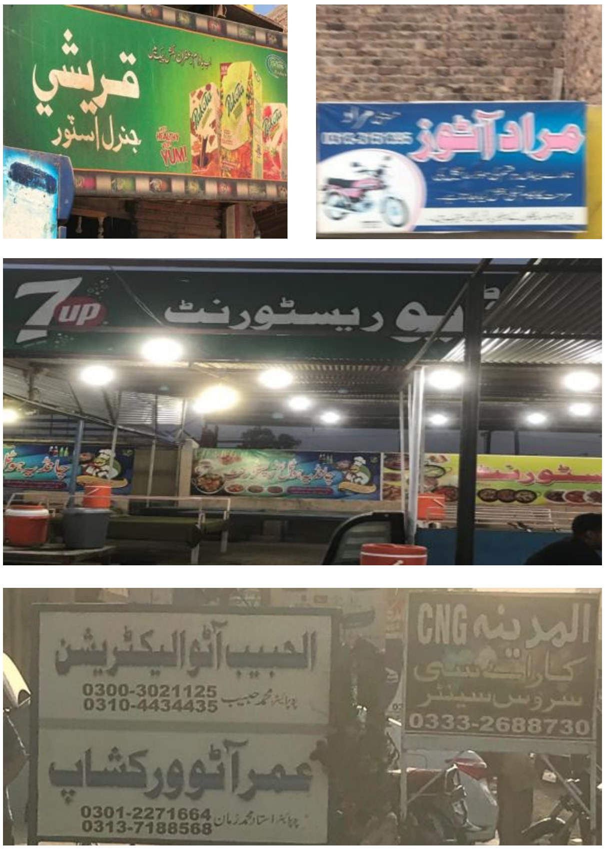


FIG. 5. SAMPLES OF SOURCE IMAGES

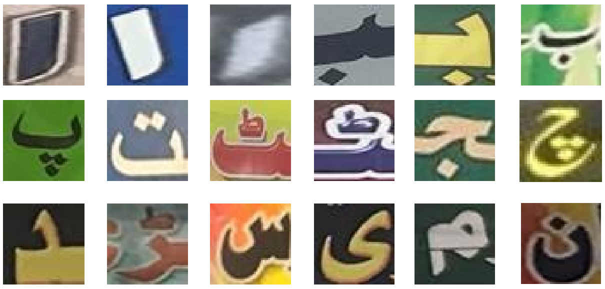


FIG. 6. SAMPLES OF NATURAL SCENE URDU CHARACTER-IMAGES

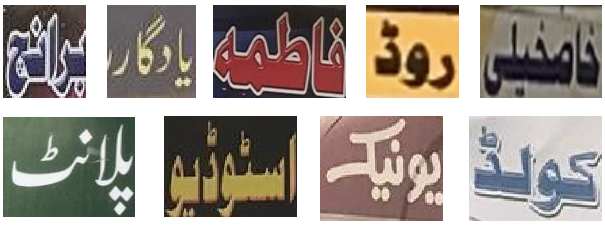


FIG. 7. SAMPLES OF URDU TEXT CROPPED WORD-IMAGES IN NATURAL SCENES

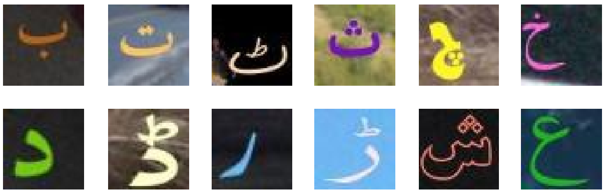


FIG. 8. SAMPLE SYNTHETIC CHARACTER-IMAGES

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, a novel dataset of Urdu text in natural scene images is described. This is the first dataset of Urdu text in natural scenes, which can be used as a standard benchmark for text detection and recognition. The proposed dataset contains 2200 images with Urdu text, 8000 cropped words and 16000 manually segmented characters. Compared to existing datasets of English, Arabic and other languages in natural scene text, the proposed dataset contains more number of images, number of cropped words characters. The segmented characters have many variations and are in different shapes (initial, media, final and isolated) within the words. Some characters are not frequently used in the text, while others are more commonly used. Therefore, the numbers of samples of each character class are not equal. To handle this class imbalance problem, a dataset of synthetic Urdu characters on the real background images with varying font sizes, colors, and alignments is also developed. Different algorithms can be developed to recognize the Urdu text in natural scenes and synthetic text in synthetic images. In future, the dataset will further be increased with more cropped words and characters. The ground truth bounding boxes at word level will also be created.

## ACKNOWLEDGEMENT

The authors are thankful to the University of New South Wales, Australia and Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan for providing research opportunities.

## REFERENCES

[1] Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R, "ICDAR 2003 robust reading competitions", In Proceedings of IEEE 7th International Conference on Document Analysis and Recognition, IEEE Computer Society, pp. 682-682, 2003.

[2] Shahab A, Shafait F, Dengel A., "ICDAR 2011 Robust Reading Competition - Challenge 2: Reading Text in Scene Images", In Proceedings of IEEE 11th International Conference of Document Analysis and Recognition, IEEE CPS, pp. 1491-1496, 2011.

[3] Karatzas D, Mestre SR, Mas J, Nourbakhsh F, Roy PP., "ICDAR 2011 Robust Reading Competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email)", In Proceedings of IEEE 11th International Conference of Document Analysis and Recognition, IEEE CPS, pp. 1485-1490, 2011.

[4] Karatzas D, Shafait F, Uchida S, Iwamura M, i Bigorda LG, Mestre SR, Mas J, Mota DF, Almazan JA, De Las Heras LP. , "ICDAR 2013 Robust Reading Competition", In Proceedings of IEEE 12th International Conference of Document Analysis and Recognition, IEEE CPS, pp. 1115-1124, 2013.

[5] Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar VR, Lu S, Shafait F, "ICDAR 2015 Competition on Robust Reading", In Proceedings of IEEE 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1156-1160, 2015.

[6] Nayef N, Yin F, Bizid I, Choi H, Feng Y, Karatzas D, Luo Z, Pal U, Rigaud C, Chazalon J, Khlif W., "ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT,". In 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1454-1459, Kyoto, 2017.

[7] Wang K, Babenko B, Belongie S. End-to-end scene text recognition. In IEEE International Conference on Computer Vision (ICCV), pp. 1457-1464, Nov 6, 2011.

[8] De Campos TE, Babu BR, Varma M.. Character Recognition in Natural Images. in Proceedings of the International Conference on Computer Vision Theory and Applications. 2009.

[9] Mishra A, Alahari K, Jawahar CV. Scene text recognition using higher order language priors. In Proccedings of BMVC-British Machine Vision Conference , Sep 3 2012.

[10] Tu Z, Ma Y, Liu W, Bai X, Yao C. Detecting texts of arbitrary orientations in natural images. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1083-1090, Jun 1 2012.

[11] Tounsi M, Moalla I, Alimi AM., "ARASTI: A database for Arabic scene text recognition," In Proceedings of 1st IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR), pp. 140-144, Nancy, 2017.

[12] Chandio AA, Pickering M, Shafi K., "Character classification and recognition for Urdu texts in natural scene images," In Proceedings of International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, pp. 1-6, 2018.

[13] Chandio AA, Pickering M, Shafi K., "Urdu Natural Scene Character Recogntion using Convolutional Neural Networks". In Proceedings of 2nd IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR), London, 2018

[14] Ramana OV, Roy S, Narang V, Hanmandlu M, "Devanagari Character Recognition in the Wild", International Journal of Computer Applications, , vol. 38, No, 4, pp. 38-45, January 2012.

[15] Tian S, Bhattacharya U, Lu S, Su B, Wang Q, Wei X, Lu Y, Tan CL., "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients", Pattern Recognit., vol. 51, pp. 125-134, Mar. 2016.

[16] Ahmed SB, Naz S, Razzak MI, Yousaf R, "Deep learning based isolated Arabic scene character recognition,"In Proceedings of 1st IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR), pp. 46-51, Nancy, 2017.

[17] Siddiqi I, Raza A., "A database of artificial urdu text in video images with semi-automatic text line labeling scheme", In Proceedings of Fourth International Conferences on Advances in Multimedia (MMEDIA'12), pp. 75-80, 2012.

[18] Jaderberg M., Simonyan K., Vedaldi A., and Zisserman A.,. Synthetic data and artificial neural networks for natural scene text recognition. In Proceedings of NIPS Workshop on Deep Learning, 2014

[19] Gupta A., Vedaldi A., Zisserman A., Synthetic data for text localisation in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315-2324, 2016.

[20] Saxena A, Chung SH, Ng AY. Learning depth from single monocular images. In Advances in neural information processing systems, pp. 1161-1168, 2006.

[21] Saxena A, Sun M, Ng AY. Make3d: Learning 3d scene structure from a single still image. In IEEE transactions on pattern analysis and machine intelligence, vol. 31, No. 5, pp. 824-40, May 2009.

[22] Gould S, Fulton R, Koller D. Decomposing a scene into geometric and semantically consistent regions. In 12th IEEE International Conference on Computer Vision, pp. 1-8, Sep 2009.