Data in Brief 31 (2020) 105749

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Cursive-Text: A Comprehensive Dataset for End-to-End Urdu Text Recognition in Natural Scene Images

Asghar Ali Chandio[a,b,*], Md. Asikuzzaman[a], Mark Pickering[a], Mehwish Leghari[b,c]

[a] School of Engineering and Information Technology, University of New South Wales, Canberra, Australia
[b] Department of Information Technology, Quaid-e-Awam University of Engineering, Science and Technology, Pakistan
[c] Institute of Information and Communication Technology, University of Sindh, Pakistan

ARTICLE INFO

Article history:
Received 17 April 2020
Revised 15 May 2020
Accepted 18 May 2020
Available online 21 May 2020

Keywords:
Urdu natural scene text dataset
Urdu text detection
Urdu text recognition
Natural scene images
Cursive text in the wild
Multilingual text spotting dataset
Convolutional neural networks

ABSTRACT

Reading text in natural scene images is an active research area in the fields of computer vision and pattern recognition as text detection, text recognition and script identification are required. In this data article, a comprehensive dataset for Urdu text detection and recognition in natural scene images is presented and analysed. To develop the dataset, more than 2500 natural scene images were captured using a digital camera and a built-in mobile phone camera. Three separate datasets for isolated Urdu character images, cropped word images and end-to-end text spotting were developed. The isolated Urdu character and cropped word images dataset contain a much larger number of samples than existing Arabic natural scene text datasets. The Urdu text spotting dataset contains images with Urdu, English and Sindhi text instances. However, the focus has been given to the Urdu text instances. The ground truths for each image in the isolated character, cropped word or text spotting datasets are provided separately. The proposed datasets can be used to perform Urdu text detection and recognition or end-to-end recognition in natural scenes. These datasets can also be helpful to develop Arabic and Persian natural scene text detection and recognition systems, as Urdu is a derived

* Corresponding author(s)
E-mail addresses: a.chandio@student.adfa.edu.au, asghar.ali@quest.edu.pk (A.A. Chandio).

https://doi.org/10.1016/j.dib.2020.105749

language of these scripts and has many similar letters. The datasets can also be helpful to develop multi-language translation systems, which can facilitate foreign tourists to read and translate multilingual text in natural scene images. To evaluate the datasets, state-of-the-art machine learning and deep neural networks were used to build the text detection and recognition models, where the best classification accuracies are achieved. To the best of the authors' knowledge, this is the first dataset proposed for Urdu text detection, recognition or end-to-end text recognition in natural scene images. The aim of this data article is to present a benchmark work in the field of document analysis and recognition.

Computer Science

Computer Vision and Pattern Recognition

Tables
Figures
Images
Text Files

Using a digital camera with a 20 megapixels (MP) sensor, an iPhone with a 12 MP back camera and a Samsung mobile with a 16MP back camera.

Raw
Analyzed

Environmental factors such as illuminations, blurring and lighting conditions were considered while capturing images. The focus was given to the text within an image.

The images in the dataset were obtained from the advertisement banners, sign-boards along the road side and streets, shop name boards, text written on the passing vehicles and walls.

The images provided in this dataset were collected in different cities of Sindh, Pakistan.

Summarized data are hosted with the article.

The datasets and their related files are hosted in a Mendeley public data repository.
DOI: https://data.mendeley.com/datasets/k5fz57zd9z/1
URL: http://dx.doi.org/10.17632/k5fz57zd9z.1

Specifications Table

Subject	Computer Science
Specific Subject Area	Computer Vision and Pattern Recognition
Type of Data	Tables Figures Images Text Files
How Data were Acquired	Using a digital camera with a 20 megapixels (MP) sensor, an iPhone with a 12 MP back camera and a Samsung mobile with a 16MP back camera.
Data Format	Raw Analyzed
Parameters for Data Collection	Environmental factors such as illuminations, blurring and lighting conditions were considered while capturing images. The focus was given to the text within an image.
Description of Data Collection	The images in the dataset were obtained from the advertisement banners, sign-boards along the road side and streets, shop name boards, text written on the passing vehicles and walls.
Data Source Location	The images provided in this dataset were collected in different cities of Sindh, Pakistan.
Data Accessibility	Summarized data are hosted with the article. The datasets and their related files are hosted in a Mendeley public data repository. DOI: https://data.mendeley.com/datasets/k5fz57zd9z/1 URL: http://dx.doi.org/10.17632/k5fz57zd9z.1

Value of the Data

- The Urdu-Text datasets are the first to be published for Urdu text detection and recognition in natural scene images.
- The datasets will be useful to help research community of computer vision and pattern recognition to develop multi-language text detection and recognition systems.
- The currently available datasets for cursive scripts such as Arabic and Persian contain a small number of images. These datasets will help to further standardize the evaluation of Arabic and Persian text in natural scene images.
- The datasets will help to develop real-world applications for camera-captured images such as content-based image retrieval, natural scene text to speech conversion, intelligent transportation systems and many more.
- The datasets contain a huge variety of images in terms of text size, writing styles, aspect ratios, background complexities and handwritten text.
- The datasets are more comprehensive than other cursive text datasets, which will facilitate new research for cursive text detection and recognition in natural images.

1. Data

1.1. Data Collection

More than 2500 images of signboards along the roadsides, street name boards, shop name boards, advertisement banners, text written on the passing vehicles and building walls were captured. These images were captured in different parts of the Sindh, Pakistan from a digital camera with a 20 MP sensor, an iPhone with a 12 MP back camera and a Samsung mobile with a 16 MP back camera. The high-resolution camera sensors were used to capture images with small text and under bright lighting conditions. All the images were taken in an uncontrolled environment in the daytime. Though the weather and environmental factors were considered while photographing images, some images with no text instances, blur and uneven lighting were discarded. Text in the images appears with multiple variations in font size, aspect ratios, text colors, background complexities, context-sensitivity and writing styles. Some examples of the photographed images are shown in Fig. 1. Besides text, natural scene images have several other objects, which make the text detection and recognition problem more complex. In addition, text in the photographed images varies from the handwritten on building walls to the machine printed signboards as illustrated in Fig. 2.

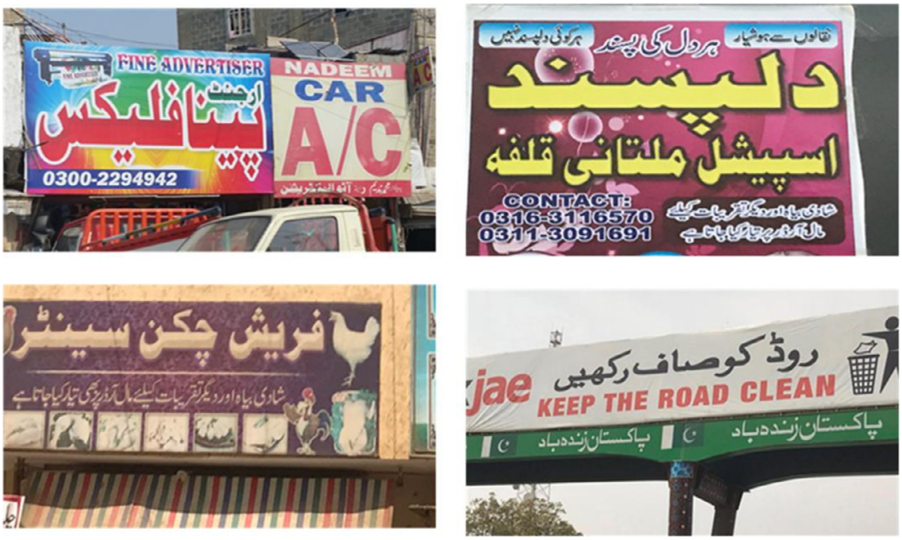


Fig. 1. Natural scene images with text variations and complex backgrounds

Detection and recognition of such handwritten text in natural images is more complex and challenging due to variations in writing styles, placing text on top of other text, and other complexities. The images captured for the dataset contain Urdu, English, Sindhi and some Arabic text instances. In Asian subcontinent countries such as Pakistan, English is used as the main language along with other local languages i.e., Urdu and Sindhi. Therefore, a large number of scene images in the Urdu-Text datasets are embedded with bilingual and some with tri-lingual text, as shown in Fig. 3. Due to the presence of multilingual text in natural images, it is necessary to develop scene text detection and recognition systems which can support multilingual scripts.

1.2. Dataset Complexities and Challenges

Detecting and recognizing cursive text in natural scene images has several unique challenges that make it more complex and challenging than applications for non-cursive text such as English. The proposed datasets contain several images with challenging and complex text. The following subsections explain some challenges associated with Urdu text in natural scene images.

1.2.1. Ligature Overlapping and Context Sensitivity

Similar to Arabic and Persian scripts, Urdu is a bidirectional cursive (connected) script, where the text is written from right to left, and the numbers are written from left to right. It is the national language of Pakistan and is widely spoken in southern Asia [1]. The Urdu alphabet contains 40 letters with 39 basic letters, whereas the Arabic alphabet has 28 and the Persian alphabet 32 letters respectively. The alphabet is derived from Persian and Arabic scripts, but it includes some additional letters, which are not present in the Persian or Arabic scripts. These scripts use almost similar alphabetical characters but differ in the number used as shown in Table 1. Urdu script differs from Arabic script in its writing style, i.e., Urdu is more commonly written in Nastaliq style while Arabic text is written in Naskh style. Characters in the same word are connected through a baseline with their predecessor and successor. The main challenges offered by cursive scripts are the non-uniform inter ligature overlaps (between the letters of two ligatures), intra-ligature overlaps (between the letters within the same ligature), ligature overlapping on text written below or above the baseline, the diagonal shape of the characters, context-

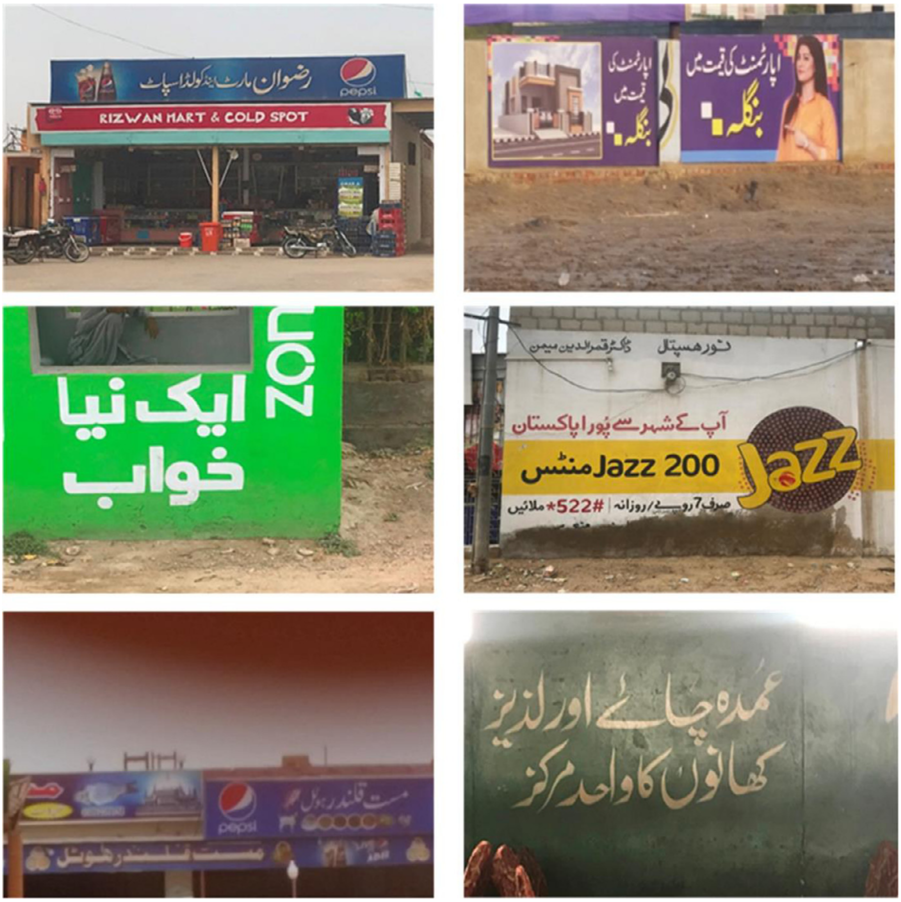


Fig. 2. Some examples of scene images in the Urdu-Text dataset. first row: machine printed text, second row: handwritten text, third row: handwritten and machine printed text with blur and uneven lighting conditions.

Table 1
Characteristics of Urdu, Arabic and Persian Scripts

Characteristics	Urdu	Arabic	Persian
Number of Characters	39	28	32
Writing Direction	Right to left	Right to left	Right to left
Cursive	Yes	Yes	Yes
Dots and Diacritics	Yes	Yes	Yes

sensitivity, and variations in the shapes of the same letter. Examples of these complexities are shown in Fig. 4.

1.2.2. Joiner and Non-joiner Letters

The Urdu-Char dataset contains images of joiner and non-joiner letters. The joiner letters change their shape depending upon the neighbouring letter or their position within a word. Hence, all the joiner letters have four basic shapes (initial, medial, final or isolated). On the other hand, the non-joiner letters can only appear in either isolated or final shape. When a word ends with a joiner letter, it is mandatory to insert space as a separator, otherwise it will

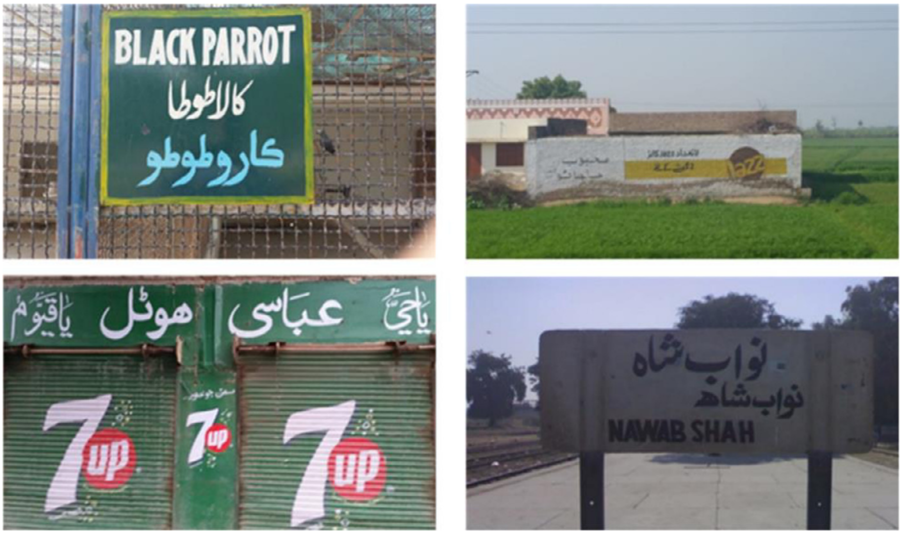


Fig. 3. Some example images in the dataset with multilingual text.

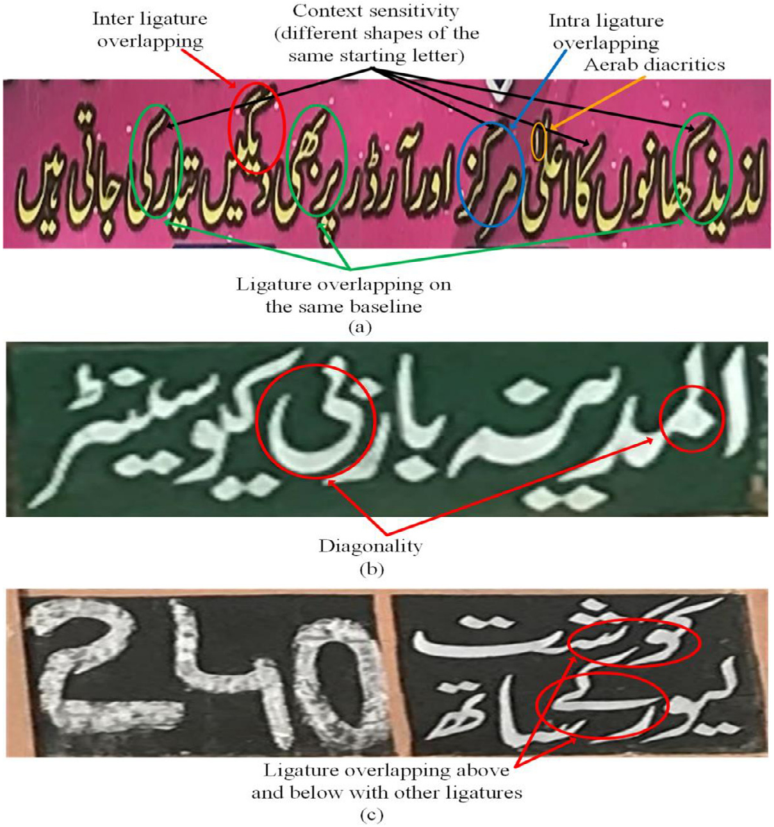

Inter ligature overlapping
Context sensitivity (different shapes of the same starting letter)
Intra ligature overlapping
Aerab diacritics
Ligature overlapping on the same baseline
(a)
Diagonality
(b)
240
Ligature overlapping above and below with other ligatures
(c)

Fig. 4. Some examples with challenging text in the Urdu-Text dataset. (a) ligature overlapping and context-sensitivity (b) diagonal text and (c) ligature overlapping on different baselines

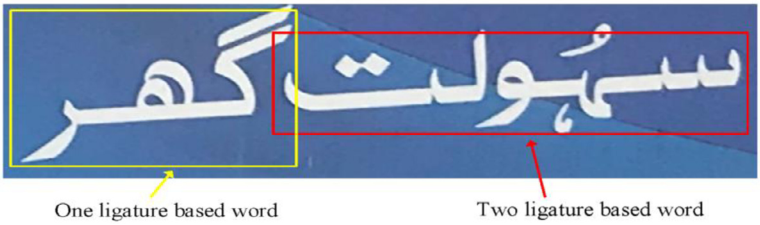


Fig. 5. Representation of ligatures in an Urdu word

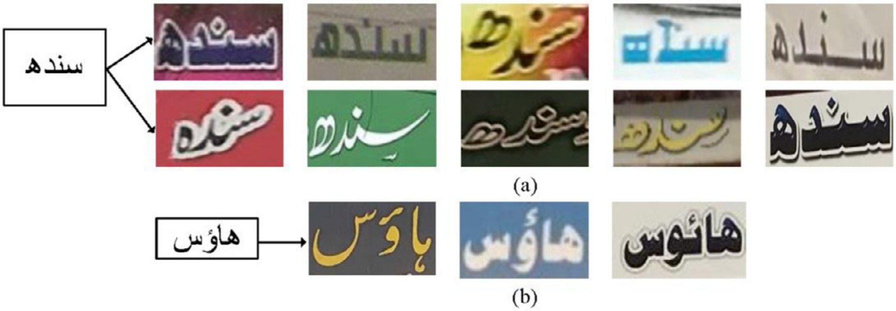


Fig. 6. Variations in writing styles. (a) the Urdu word ھدنس written in ten different styles (b) placing a ligature on top of another ligature

merge the current word with the following letter, which has no meaning in the Urdu script. For example, two different words لابقا یرٹیب without inserting a space will become یرٹیبلابقا, which makes the word incorrect. However, the non-joiner letters do not need a space inserted before the next letter, which is, therefore, a ligature of the same word. For example, the word لوکسا looks like three separate words, but because the ligatures do not have any space between them, they are three ligatures ا, وکس and ل of the same word. In cursive scripts like Urdu, two or more letters are combined according to the joiner or non-joiner property to form a unique shape called a ligature or a sub-word. A word is then formed by combining one or more ligatures as shown in Fig. 5.

1.2.3. Dots and Diacritics

In the Urdu-Text datasets, several letters are surrounded by one or more special marks called a diacritics, which are placed above or below the letters. Three types of diacritics can be used in the Urdu script: dots, aerabs and a superscript ط. The dots are placed either above, below or within the associated letters. The number of dots varies from one to a maximum of three. Several letters within the same character groups are differentiated either by the number of dots or their positions. The Urdu alphabet has 17 letters, which are accompanied by dots. The aerab diacritics are optional and sometimes used to eliminate pronunciation errors. These are placed above or below the letters such as تُ, ت, and تٰ. The superscript ط is placed only above three letters in the Urdu text. These letters are called retro-flex constants and are only present in the Urdu script.

1.2.4. Variations in the Writing Styles

Urdu is a type of cursive script, where a single letter can have up to four shapes. Every character changes its shape according to the neighbouring letter. In addition, the same letter or word can be written with various writing styles. This makes the scene text recognition problem more challenging. As shown in Fig. 6(a), a single natural scene Urdu word ھدنس in the cropped word image dataset is written with ten different writing styles by varying either the stroke width of

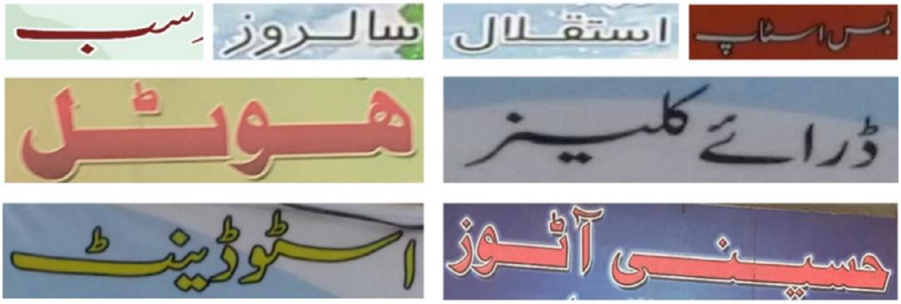


Fig. 7. Some examples of stretched text in the cropped Urdu word image dataset

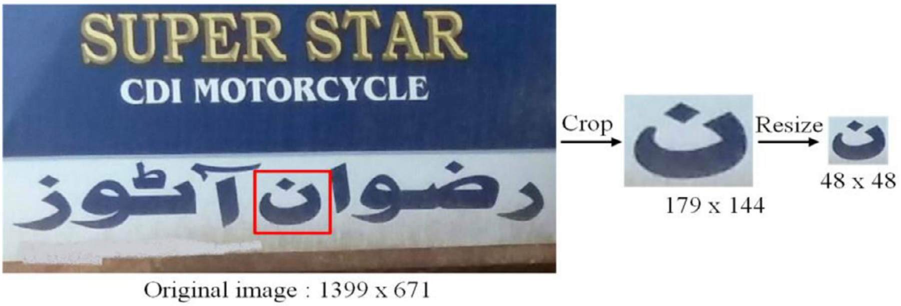


Fig. 8. Manual segmentation of characters

the letters or their styles. In the Urdu script, it is common to place a ligature on top of another ligature in the same word. This technique, called positioning, is used to accommodate long text within a small space. In Fig. 6(b), the ligature ء is placed on top of the ligature و. Moreover, the ligature ء, as shown in Fig. 6(b), is written with different writing styles. All the ligatures in Fig. 6(b) represent the same word.

In addition to variations in writing styles, several images in the dataset contain horizontally stretched letters or words, where they completely change their standard shape into a wider shape, which increases the width of the letters or a complete word. The stretching property of text in natural scenes may make it difficult to detect exactly the same text as in the ground truth. In addition, it is more difficult to recognize stretched text than text in its standard shape. Fig. 7 shows some examples of stretched text.

1.3. Description of the Datasets

In this paper, three separate datasets have been developed for isolated Urdu character recognition, cropped Urdu word recognition, and multilingual text detection and recognition. Following subsections briefly describe each dataset.

1.3.1. Character Image Dataset

The Urdu-Char dataset contains 19901 character images. All the character images were manually segmented and annotated. The process of character segmentation is shown in Fig. 8. The Urdu-Char dataset represents character images with their different forms (isolated, initial, medial and final).

Some of the Urdu characters such as ث, ظ and ژ are not frequently used; therefore, the Urdu-Char dataset contains an unequal number of images for each of the character classes. The liga-

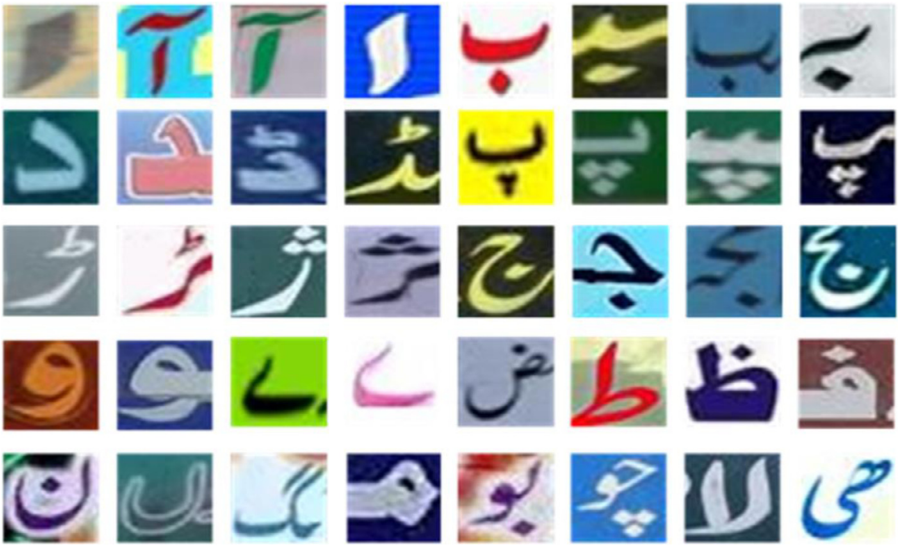


Fig. 9. Some examples of isolated character images in Urdu-Char dataset

Table 2
Comparison of the Urdu-Char and Urdu-Word datasets with related datasets

Datasets	No. of Images	No. of Character Images	No. of Word Images
ICDAR03 [4]	251	6185	1157
ICDAR13 [5]	462	–	5003
ICDAR15 [6]	1670	–	6545
Chars74K [7]	1922	7705 English, 3345 Kannada	1416
ICDAR17 MLT Arabic [8]	800	–	3712
ARASTI [2]	371	2093	1687
EASTR [3]	2469	16624 Arabic, 5904 English	2593 Arabic, 5172 English
Urdu-Text	**2500**	**19901**	**14100**

tures such as اط, اک and others, which are formed by connecting only two letters, are considered to be a single character. Some examples of the Urdu-Char image dataset with their different forms are shown in Fig. 9, and a comparative analysis of the Urdu-Char dataset with English, Kannada and Arabic datasets is shown in Table 2.

1.3.2. Word Image Dataset

The Urdu-Word dataset contains 14100 cropped images of the Urdu text with their annotations in a UTF-8 encoded text file. All the words were manually segmented and resized to a fixed width and height of 100 × 64 × 3. This cropped Urdu-Word image dataset is used for text recognition purposes. A lexicon of 40000 commonly used Urdu words was also created as a separate text file that will be useful to address word recognition errors. The cropped dataset contains mainly images with Urdu text, some with Sindhi text and digits. The word images with English text were not considered at the moment because of the existing large-scale freely available English word datasets. The Urdu-Word dataset contains scene images with different text styles, sizes, colors and backgrounds as well as un-even lighting and arbitrary oriented text. Compared to the Arabic word image datasets developed in [2] and [3], the Urdu-Word dataset contains a larger number of images. Therefore, it can be used as a benchmark dataset for word spotting problems. A comparison of the cropped Urdu-Word image dataset with Englihs, Kannada and Arabic datasets is presented in Table 2. Some examples from the cropped Urdu-Word image dataset are shown in Fig. 10.

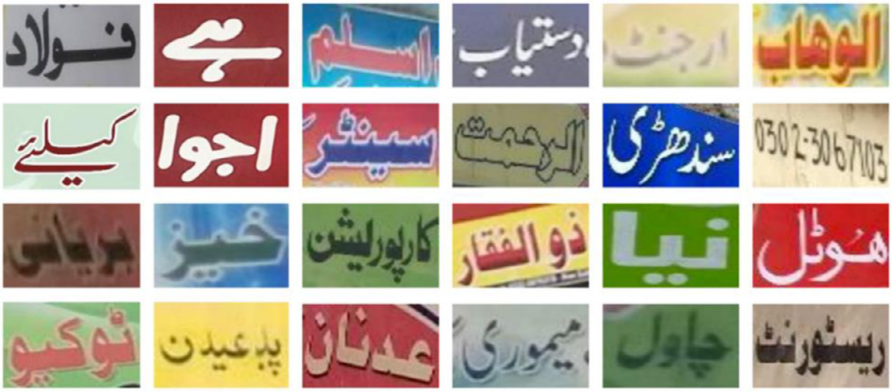
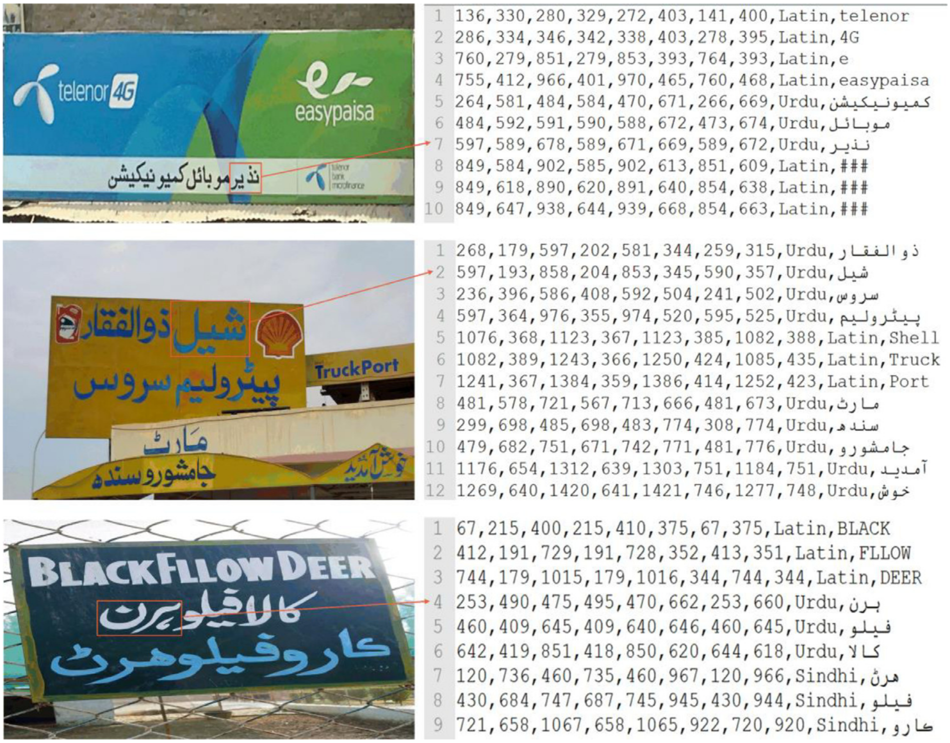


Fig. 10. Some examples of cropped Urdu word images.



```
136,330,280,329,272,403,141,400,Latin,telenor
286,334,346,342,338,403,278,395,Latin,4G
760,279,851,279,853,393,764,393,Latin,e
755,412,966,401,970,465,760,468,Latin,easypaisa
264,581,484,584,470,671,266,669,Urdu,کمیونیکیشن
484,592,591,590,588,672,473,674,Urdu,موبائل
597,589,678,589,671,669,589,672,Urdu,نذیر
849,584,902,585,902,613,851,609,Latin,###
849,618,890,620,891,640,854,638,Latin,###
849,647,938,644,939,668,854,663,Latin,###
```

```
268,179,597,202,581,344,259,315,Urdu,ذوالفقار
597,193,858,204,853,345,590,357,Urdu,شیل
236,396,586,408,592,504,241,502,Urdu,سروس
597,364,976,355,974,520,595,525,Urdu,پیٹرولیم
1076,368,1123,367,1123,385,1082,388,Latin,Shell
1082,389,1243,366,1250,424,1085,435,Latin,Truck
1241,367,1384,359,1386,414,1252,423,Latin,Port
481,578,721,567,713,666,481,673,Urdu,مارٹ
299,698,485,698,483,774,308,774,Urdu,سندھ
479,682,751,671,742,771,481,776,Urdu,جامشورو
1176,654,1312,639,1303,751,1184,751,Urdu,آمدید
1269,640,1420,641,1421,746,1277,748,Urdu,خوش
```

```
67,215,400,215,410,375,67,375,Latin,BLACK
412,191,729,191,728,352,413,351,Latin,FLLOW
744,179,1015,179,1016,344,744,344,Latin,DEER
253,490,475,495,470,662,253,660,Urdu,برن
460,409,645,409,640,646,460,645,Urdu,فیلو
642,419,851,418,850,620,644,618,Urdu,کالا
120,736,460,735,460,967,120,966,Sindhi,هرٹ
430,684,747,687,745,945,430,944,Sindhi,فیلو
721,658,1067,658,1065,922,720,920,Sindhi,ڪارو
```

Fig. 11. Some sample images in Urdu-Text spotting dataset. left: Urdu-Text images, right: annotation files.

1.3.3. Text Spotting Dataset

The Urdu-Text spotting dataset comprises 1400 images with all the text instances manually annotated with rectangular bounding boxes. Fig. 11 illustrates some examples from the Urdu-Text spotting dataset with their annotation files. As mentioned in Section 1, a large number of images in the dataset contain bi-lingual or tri-lingual text. The statistics of the word instances for each script in the dataset is given in Table 3. The Urdu-Text spotting dataset is split into 1000 training and 400 testing images. This dataset has 9.85 average text instances per image, while

Table 3
Statistics of the Word Instances for each script

Script Type	No. of Words
Urdu	7603
English	5653
Sindhi	350
Arabic	68
Symbols	113
Others	01

the ICDAR15 [6], ICDAR17 MLT Arabic [8], COCO-Text [9] and Total-Text [10] datasets have 7.12, 9.34, 2.73 and 7.37 average text instances per image, respectively. The Urdu-Text spotting dataset may be used with ICDAR17 MLT or ICDAR19 MLT datasets to solve the problems of multilingual text detection, text recognition or script identification.

2. Experimental Design, Materials and Methods

To evaluate the Urdu-Text dataset, separate experiments were performed for isolated character recognition, cropped word recognition and multilingual text detection with three datasets, which are discussed in the following subsections.

2.1. Isolated Urdu Character Recognition

The Urdu-Char dataset was split into training (70%) and test (30%) sets. First, the input images were re-scaled to a fixed size of $48 \times 48 \times 3$ and then converted into grey-scale. In the experiments, Histogram of Oriented Gradients (HOG) [11] and Local Binary Pattern (LBP) [12] features were separately extracted from the training and test sets and then combined to recognize the isolated Urdu characters. The combination of two different feature descriptors created more powerful features. A linear support vector machine (SVM) classifier was trained on the combined features. The classification accuracy of the system was measured in terms of precision, recall and F-score. Precision measures the number of true character classes predicted by the classifier that originally belong to the ground truths. Recall measures the number of true character classes predicted by the classifier out of all the true ground truths in the dataset. F-Score is a weighted average of both precision and recall. Table 4 shows the detailed classification accuracy of each isolated Urdu character class. As shown in Table 4, the classification accuracy of several characters which have a similar shape such as ب, ت, ج, غ and ح is lower than for the letters with unique shapes.

2.2. Cropped Word Recognition

To eliminate the problem of character segmentation, a segmentation free method was used to recognise text from a word image. A convolutional neural network (CNN), similar to VGG-16 [13], was used to extract the features, which were encoded into sequences by a recurrent neural network (RNN) without pre-segmenting the characters. A connectionist temporal classification (CTC) method was used to decode the output of the RNN into target text transcriptions. Characters in cursive script are often stretched horizontally and occupy more width, as illustrated in Fig. 7. In this case, one character may occupy several time-steps and fall into more than one receptive field. In such cases, transforming the output of each time-step with an RNN is likely to produce more incorrect classification results. To solve this problem, a network with a CTC loss function is trained without specifying the starting and ending positions and width (how

Table 4
Classification accuracy of each of the character class. The rows in table are ordered according to the ascending F-Score values.

Character Class	Precision	Recall	F-Score	Character Class	Precision	Recall	F-Score
غ	0.56	0.55	0.56	ذ	0.79	0.86	0.83
ع	0.62	0.55	0.58	ہ	0.89	0.79	0.84
چ	0.60	0.56	0.58	ھ	0.88	0.81	0.85
ت	0.63	0.64	0.63	ـہ	0.83	0.88	0.85
ح	0.65	0.61	0.63	ـی	0.85	0.86	0.86
خ	0.65	0.60	0.63	ط	0.85	0.87	0.86
ب	0.64	0.65	0.65	م	0.87	0.85	0.86
ض	0.59	0.77	0.67	ز	0.85	0.92	0.88
پ	0.69	0.68	0.68	ظ	0.89	0.87	0.88
ص	0.68	0.70	0.69	ڈ	0.88	0.88	0.88
ف	0.70	0.70	0.70	س	0.88	0.90	0.89
ث	0.71	0.70	0.71	ٹ	0.91	0.87	0.89
ج	0.71	0.71	0.71	ک	0.90	0.91	0.90
ء	0.67	0.77	0.72	ل	0.92	0.88	0.90
ں	0.81	0.68	0.74	ے	0.91	0.93	0.92
ش	0.77	0.72	0.74	ا	0.92	0.96	0.94
ق	0.84	0.66	0.74	و	0.95	0.94	0.94
ن	0.76	0.75	0.75	ی	0.95	0.95	0.95
گ	0.76	0.76	0.76	ر	0.94	0.96	0.95
ڑ	0.88	0.70	0.78	ژ	0.98	0.93	0.95
د	0.78	0.87	0.82	آ	0.94	0.97	0.96

Table 5
Accuracy for Urdu word image text recognition

Model	RNN Type	No. of Hidden Units	WRR (%)
CNN + RNN + CTC	LSTM	128	74.77
CNN + RNN + CTC	BLSTM	128	78.13
Tesseract-OCR [13]	–	–	6.81

much space it occupies) of the character. A CTC function labels the unsegmented data sequences without pre-segmenting during network training and does not require post-processing methods to merge the individually recognized characters into the target words. The architecture of the CNN network consisted of seven convolutional layers, each followed by a non-linear rectification function. The input to the network was a fixed word image of 100 × 64 × 3 pixels and the output of the network was 1 × 25 x C, where 1 and 25 were the height and width of the feature map, while C represented the number of convolutional filters. To reduce the spatial size of the convolutional feature maps, six max-pooling layers were used. First, two max-pooling layers used a pool size of 2 × 2 and a stride of 2 × 2, while the remaining four layers used a pool size of 2 × 1 and a stride of 2 × 1 to down-sample only the height of the feature map while keeping the same width. The network was trained by using an Adam optimizer [15] with an initial learning rate of 0.01, which was decreased by 0.01 after every 9000 iterations. The network was trained for 30000 iterations. The error differential in the RNN was calculated by using back-propagation through time (BPTP) [16]. The network was implemented in the Tensorflow deep learning framework and the training was performed on an Intel(R) 2.30 GHz CPU with 16 GB random access memory (RAM) and an NVIDIA GeForce GTX 1050 Ti GPU. In the experiments, both the long short-term memory (LSTM) [17] and the bi-directional long short-term memory (BLSTM) [18] types of RNN were trialled and the best accuracy was achieved with the BLSTM network. The performance of the models was measured in terms of the word recognition rate (WRR) and compared with a popular multilingual commercial OCR algorithm [19]. Table 5 shows a comparison between the Urdu word recognition models and the Tesseract-OCR algorithm.

Table 6
Comparison of text detection accuracy on Urdu-Text spotting dataset

Model	Precision	Recall	F-Score
EAST [14]	0.18	0.39	0.26
CTPN [15]	0.32	0.71	0.43
Proposed CNN without Pre-trained Weights	0.22	0.45	0.30
Proposed CNN with Pre-trained Weights	0.29	0.70	0.37

2.3. Multilingual Text Detection

The Urdu-Text dataset contains multilingual text instances and was annotated at word-level. It was evaluated using three deep convolutional neural network models: (i) efficient and accurate scene Text [20], (ii) connectionist text proposal network [21] and (iii) a network based on the VGG16 [13] network with and without pre-trained weights. In the VGG16 CNN model, an addition operation was performed to add the convolutional features with the same spatial dimensions. A Fast-RCNN [22] network was used for classifying text/non-text proposals and bounding box regression. A recurrent network was built within the convolutional network to connect the text proposals sequentially. The model was trained using a stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and an initial learning rate of 0.001. A learning rate scheduler method was used to decrease the learning rate by 0.1 after every 10K iterations. The model was trained for 35K iterations. The performance of the models were measured in terms of precision, recall and F-score and the results are shown in Table 6. These results show that the deep learning models perform poorly on the Urdu-text dataset. This indicates the dataset is complex and contains natural scene images with challenging text.

Code Availability

The code and all the models trained can be provided by sending an email to the corresponding author.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The first author is thankful to the University of New South Wales, Australia for supporting his Ph.D. candidature with a scholarship.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.105749.

References

[1] D.M. Eberhard, G.F. Simons, Ethnologue: Languages of Asia, in: C.D. Fennig (Ed.), SIL International, SIL International, Dallas, Texas, 2019 Online version http://www.ethnologue.com .

[2] M. Tounsi, I. Moalla, A.M. Alimi, ARASTI: A database for Arabic scene text recognition, in: proceedings of 1st IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR), 2017, pp. 140–144.

[3] S.B. Ahmed, S. Naz, I.M. Razzak, R.B. Yusof, A novel dataset for English-Arabic scene text recognition (EASTR)-42K and its evaluation using invariant feature extraction on detected extremal regions, in: IEEE access, vol. 7, 2019, pp. 19801–19820. https://doi.org/10.1109/ACCESS.2019.2895876.

[4] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, ICDAR 2003 robust reading competitions, in: Proceedings of 7th IEEE International Conference on Document Analysis and Recognition, pp. 682-687, 2003.

[5] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L.G.I. Bigorda, S.R. Mestre, J. Mas, D.F. Mota, J.A. Almazàn, L.P.D.L. Heras, ICDAR 2013 robust reading competition, in: Proceedings of 12th IEEE International Conference on Document Analysis and Recognition, 2013, pp. 1484–1493.

[6] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, F. Shafait, ICDAR 2015 competition on robust reading, in: Proceedings of 13th IEEE International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1156–1160.

[7] T.D. Campos, B.R. Babu, M. Varma, Character Recognition in Natural Images, in: Proceedings of International Conference on Computer Vision Theory and Applications, 2009, pp. 273–280.

[8] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, W. Khlif, Icdar2017 robust reading challenge on multilingual scene text detection and script identification-rrc-mlt, in: Proceedings of 14th IEEE IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 1454–1459.

[9] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, Coco-text: Dataset and benchmark for text detection and recognition in natural images, in: arXiv, 2016.

[10] C.K. Ch'ng, C.S. Chan, Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition, in: Proceedings of 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, 2017, pp. 935–942.

[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, 2005, pp. 886–893.

[12] T. Ojala, M. Pietikinen, D. Harwood, A comparative study of texture measures with classification based on feature distributions, Pattern recognition 29 (1) (1996) 51–59.

[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition,, in: arXiv preprint arXiv:1409.1556, 2014.

[14] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[15] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: arXiv preprint arXiv:1412.6980, 2014.

[16] P.J. Werbos, Backpropagation through time: what it does and how to do it, in: Proceedings of the IEEE, vol. 78, 10, 1990, pp. 1550–1560.

[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[18] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural networks 18 (5-6) (2005) 602–610.

[19] Tesseract Open Source OCR Engine. https://github.com/tesseract-ocr/tesseract, 2020 (Accessed 9 March 2020)

[20] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, EAST: an efficient and accurate scene text detector, in: Proceedings of CVPR, July 2017, pp. 2642–2651.

[21] Z. Tian, W. Huang, T. He, P. He, Y. Qiao, Detecting text in natural image with connectionist text proposal network, in: Proceedings of Springer European conference on computer vision, 2016, pp. 56–72.

[22] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.