# Gurmukhi Text: A Dataset for Natural Scene Gurmukhi Text Detection and Recognition

**Jaspreet Kaur[1], Dr. Dharam Veer Sharma[2]**
*[1]Research Scholar, Department of Computer Science, Punjabi University, Patiala*
*[2]Professor, Department of Computer Science, Punjabi University, Patiala*

***Abstract***
Digitization of text plays a vital role in the area of image processing and pattern recognition. However, recognizing text from natural scene has become a challenging task for the researchers due to the challenges of the natural scene along with the lack of benchmark datasets. Public datasets are available for the Latin and Arabic scripts which are useful for research in the field of natural scene text recognition. But for Gurmukhi script no public dataset is available for natural scene text. This paper introduces a new dataset for the natural scene Gurmukhi text images. Dataset contains 500 images of Gurmukhi text which can be used to test the system against different challenges.  Dataset contain complete scene images and can be used for script identification, detection and recognition for Gurmukhi. This paper also provides a survey of benchmark public datasets available for natural scene text recognition.

***Keywords:*** *recognition, Gurmukhi, identification, benchmark*

## Introduction

Natural scene text recognition portrays the computer's ability to digitize text which further can be processed for various applications such as geo-localization, label recognition, to facilitate tourists through translation in the global language. Many benchmark datasets such as ICDAR 03, 05, 11, 13 [1], incidental Scene Text' [2], SVT [3], NEOCR [4], MSDR-500[5] have been published which played major role for research in Latin script recognition. However, very negligible work is done for non-Latin scripts especially for Gurmukhi script in case of natural scene text recognition as no benchmark dataset is available. Major role of this research is provide a public dataset of Gurmukhi text images which can be used for testing of existing techniques along with development of new ones.

## Gurmukhi Script

Gurmukhi script is used for writing Punjabi language which is very popular language in India and Pakistan. Around 100,000,000 people around the world use Punjabi language and this is the tenth most widely spoken language around the world. Gurmukhi follows cursive structures and written from left to right direction. There are some distinctive features of Gurmukhi script as a) It uses three zones ie. upper, middle and lower for its text formation. b) there is no uppercase or  lowercase letters in Gurmukhi. The key contribution of our work is a development of a dataset for Gurmukhi which will help the researchers to contribute in this field.

## Dataset Collection

As no benchmark dataset exists for natural scene Gurmukhi textual images, therefore we have collected experimental dataset of natural scene images using different smart phones having minimum 8 megapixel camera. Images of signboards, banners, pamphlets, hoarding, and text written on walls, book covers, charts or other natural scene containing text are captured. Our Dataset consists of 500 images of natural scene images that are available in our surrounding such as schools, gurdwaras, street signboards, banners, flex boards, bus-stands, storefronts and also from book covers that are complex in nature. Images are captured in different lightening conditions as taken from both indoor and outdoor scenes. Our dataset covers a broad range of aspects which distinguishes real world images from scanned images.  Dataset is enriched with images of different alignments and orientations, noise, blurring, uneven lightening, shadows as well as different font size and style, caused by natural scene and mobile camera. However, natural scene images contain both handwritten and machine printed text which makes text detection and recognition process more cumbersome due to different writing styles and fonts.  Images captured from streets, which consist of a large variety of complicated real-world scenarios, e.g., store fronts and landmarks, making the challenge extreme high by narrowing gaps between research and real applications. Dataset contains images based on the conditions and challenges related with textual image which are: normal, low light, shadow, blurring, noise, curved, multi oriented, skewed, perspective distortion, and cluttered background. Therefore, Dataset can be used as a benchmark to test the performance of GurmukhiOCR against different challenges.

Sample images from the collected dataset are shown in the Figure 1 that show various challenges related with complexity of natural scene images.

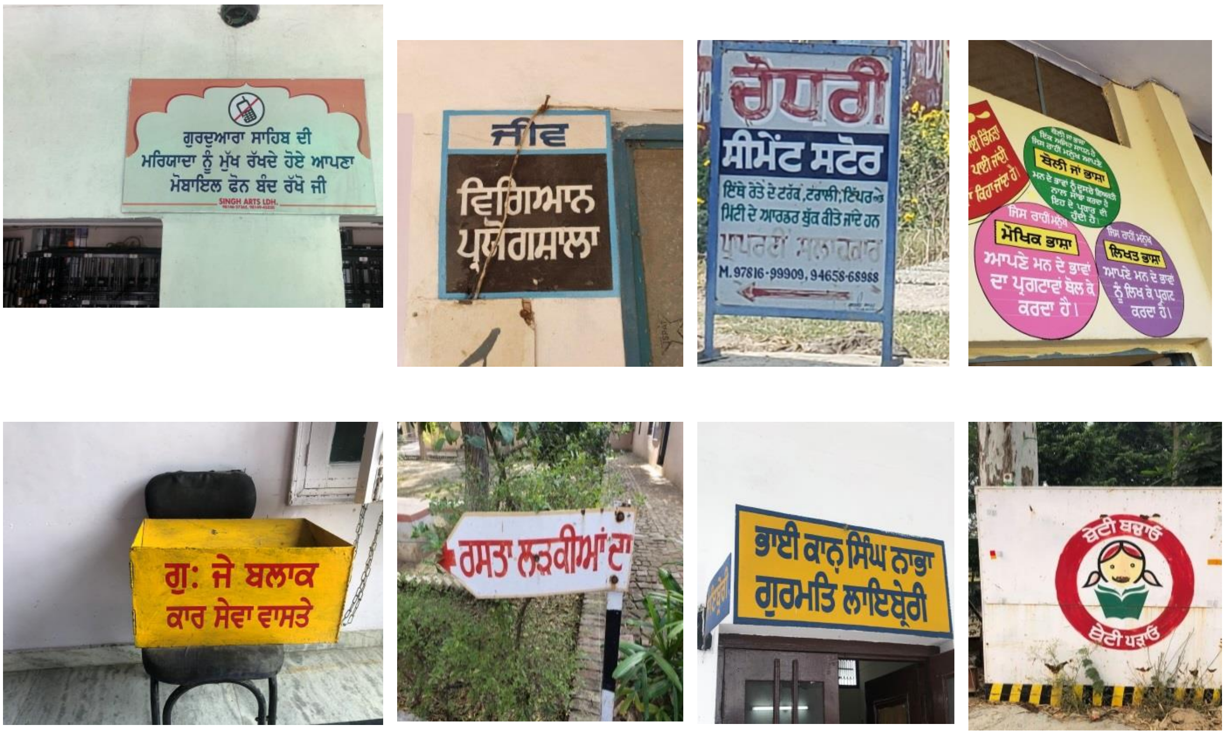

**Figure 1: Dataset Sample Images**

**Literature Survey on benchmark Datasets for Natural Scene Text Detection**

Several benchmark datasets are available for detection of text from natural scene images for both Latin and non-Latin scripts. These benchmark algorithms can be used to evaluate text detection algorithms for natural scene images.

**ICDAR [1] has** three variants. ICDAR’ 03 started out with 509 camera taken scene text images. All the scene texts in the dataset appear in horizontal orientation. In ICDAR’11 , the total number of images were reduced to 484 to eliminate duplication in the previous version.ICDAR’13 further trimmed down the 2011 version to 462 images of horizontal English text only, improvement was done to increase its text categories and tasks. Recently, ICDAR launched a new challenge named as the ‘incidental Scene Text’ **[2]** which is based on 1670 images captured with wearable devices. It is more challenging than previous datasets as it has included text with arbitrary orientation and most of them are out of focus.

**Street View Dataset (SVT) [3]** contains 350 pictures captured from Google Street View and annotated at word level using Axis aligned bounding boxing. This dataset has not annotated all instances of text.

**NEOCR: Natural Environment OCR Dataset [4]** consists of 659 real world images with 5238 annotated bounding boxes (text fields). Dataset is collected by different persons independently so contains broad range of characteristics which distinguish natural scene images from document text. All text recognizable by humans has been annotated for all images.

**MSRA Text Detection 500 Database [5]** is collected and released publically as a benchmark to evaluate text detection algorithms for Chinese and English scripts. Images are taken from both indoor and outdoor scenes such as offices, malls and streets which are mainly focused on billboards, guideboards, caution plates etc. Dataset is challenging due to the presence of complex backgrounds, multi-orientations, color and size of text. Dataset contain total 500 natural scene images from which any random 300 can be used as training set and other 200 for testing set.

**CUTE80 [6]** dataset is created with 80 complex background, curved text line images having low resolution and perspective distortion issues. This dataset is a benchmark to test the performance of a method against curved and low resolution text images of English.

**COCO-Text Dataset [7]** is the largest dataset available with 63686 annotated images and 173589 text instances. The source of images is Microsoft COCO dataset. Images are having versatile properties such as blurriness, clarity and both handwritten and machine printed.

**Total-Text Dataset [8]** is a comprehensive scene text dataset which contains 1555 images which are collected from internet sources and captured from real world scenes that put emphasis on curved text images which is a missing feature in most of the available datasets as those focus on horizontal and multi-oriented text. Total-Text is created with diversified orientation images and almost fifty percent pictures contain more than two orientations. It is annotated for three different tasks: detection, recognition, and segmentation. English is the only annotated language; the rest of the language was labelled as 'do not care' region.

**RCTW17 [9]** dataset is designed with more than 12000 images along with complete annotations which are created using polygon box drawing around the text. Location and transcription of text is also included in the annotations.

**CTW [10]** is a large dataset having Chinese text images which are annotated at character level. Dataset ignores other languages text present in the images and annotates only Chinese text.

**CTW1500 [11]** is created with curved text Chinese images where each image contains at least one curved text along with arbitrary orientations.

**Large Scale Street View Text with Partial Labelling (LSVT) [12]** consists of 20,000 testing data, 30,000 training data in full annotations and 400,000 training data in weak annotations, which are referred to as partial labels. For most of the training data in weak labels, only one transcription per image is provided, which is referred as `text-of-interest'.

**Table 1: A Comparison of Publically Available Natural Scene Text Datasets**

| **Dataset** | **Year** | **Language** | **Number of Images** | **Training Images** | **Testing Images** | **Text Orientation** | **Annotation** |
|---|---|---|---|---|---|---|---|
| ICDAR 2003 | 2003 | English | 509 | 258 | 251 | Horizontal | Word |
| SVT | 2010 | English | 350 | 100 | 250 | Arbitrary | Word |
| ICDAR 2011 | 2011 | English | 484 | 229 | 255 | Horizontal | Word/Characte |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | r |
| NEOCR | 2011 | English | 659 | 339 | 320 | Multi-oriented | Word |
| MSRA-TD500 | 2012 | English /Chinese | 500 | 300 | 200 | Arbitrary | Text Line |
| ICDAR 2013 | 2013 | English | 462 | 229 | 233 | Horizontal | Word/Character |
| CUTE80 | 2014 | English | 80 | - | - | Curved | Text Line |
| ICDAR 2015 | 2015 | English | 1500 | 1000 | 500 | Multi-oriented | Word |
| COCO TEXT | 2016 | English | 29210 | 22184 | 7026 | Arbitrary | Word |
| TOTAL TEXT | 2017 | English | 1555 | 1255 | 300 | Multi-oriented | Word |
| RCTW-17 | 2017 | English/Chinese | 12514 | 11514 | 1000 | Arbitrary | Text Line |
| CTW | | Chinese | 32285 | 25887 | 6398 | Arbitrary | Character/ Text Line |
| CTW1500 | 2017 | English/Chinese | 1500 | 1000 | 500 | Curved/Multi-oriented | Word/text Line |
| ArT19 | 2019 | English/Chinese | 10166 | 5603 | 4563 | Multi-oriented/ Arbitrary | Word |
| LSVT | 2019 | English /Chinese | 50000 fully annotated +400000<br>Weakly annotated | 30000 | 20000 | Multi-oriented/Arbitrary | Word/text Line |
| GurmukhiText | 2022 | Gurmukhi | 500 | 300 | 200 | Multi-oriented | Character |

## Conclusion

In the proposed paper, we have introduced a dataset for Natural scene Gurmukhi text images, with an goal to identify the existing research gap in natural scene text recognition for non-Latin scripts, specifically Gurmukhi, by introducing a new dataset of 500 natural scene Gurmukhi text images. The lack of public datasets for Gurmukhi script recognition previously hampered the progression of research in this field, limiting the advancements in applications such as geo-localization, label recognition, and language translation. The dataset developed under this research addresses different challenges in the natural scene text recognition, like varying light conditions, noise, blurring, and font styles, thus providing a comprehensive base for testing and development of new techniques.

Additionally, the paper provides a valuable survey on benchmark public datasets available for natural scene text recognition, offering insights and comparisons that can guide further research initiatives. The dataset, captured under various real-world scenarios and exhibiting the complexities inherent in natural scene images, has the potential to significantly contribute to the research community by driving advancements in text recognition algorithms and techniques, especially for Gurmukhi script.

In future works, the potential enhancement of this dataset with more diversified examples and exploring machine learning or deep learning models for improved recognition of Gurmukhi script from natural scene images would be highly beneficial. Further, the application of these advancements in practical use-cases like real-time translation or navigation assistance can be explored.

**References**

[1] D Karatzas, Shafait F, Uchida S, Iwamura M, Heras LPDL. ICDAR 2013 robust reading competition// Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE Computer Society, 2013.

[2] D Karatzas, Gomez-Bigorda L, Nicolaou A, Ghosh S, Valveny E. ICDAR 2015 competition on Robust Reading// International Conference on Document Analysis & Recognition. IEEE Computer Society, 2015.

[3] Kai Wang, B. Babenko and S. Belongie, ”End-to-end scene text recognition,” 2011 International Conference on Computer Vision, Barcelona, pp. 1457-1464, 2011.

[4] R. Nagy, A. Dicker, and K. Meyer-Wegener, "Definition and Evaluation of the NEOCR Dataset for Natural-Image Text Recognition". University of Erlangen, Dept. of Computer Science, Technical Reports, CS-2011-07, September 2011.

[5] Cong Y, Xiang B, Liu W, Yi M, Tu Z. Detecting texts of arbitrary orientations in natural images// Computer Vision & Pattern Recognition IEEE, 2012.

[6] A Robust Arbitrary Text Detection System for Natural Scene Images A. Risnumawan, P. Shivakumara, C.S. Chan and C.L. Tan Expert Systems with Applications, vol. 41(18), pp. 8027-8048, 2014.

[7] Gomez R, Shi B, Gomez L, Numann L, Karatzas D. ICDAR2017 Robust Reading Challenge on COCO-Text// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) IEEE Computer Society, 2017.

[8] Ch'Ng C K , Chan C S. Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2018.

[9] Shi B, Yao C, Liao M, Yang M, Xu P, Cui L, Belongie S, Lu S, Bai X. ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17)// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) IEEE, 2017.

[10] Yuan T , Zhe Z, Xu K, Li CJ, Hu SM. Chinese Text in the Wild. 2018.

[11] Liu Y, Jin L, Zhang S, Sheng Z. Detecting Curve Text in the Wild: New Dataset and New Solution. 2017.

[12] Yipeng Sun, Large-scale Street View Text with Partial Labeling (ICDAR-2019 LSVT) 1,ID: ICDAR-2019.