UTRNet: High-Resolution Urdu Text Recognition In Printed Documents

Abdur Rahman (✉)[0000−0002−9547−2435], Arjun Ghosh, and Chetan Arora

Indian Institute of Technology Delhi
ch7190150@iitd.ac.in

Abstract. In this paper, we propose a novel approach to address the challenges of printed Urdu text recognition using high-resolution, multi-scale semantic feature extraction. Our proposed UTRNet architecture, a hybrid CNN-RNN model, demonstrates state-of-the-art performance on benchmark datasets. To address the limitations of previous works, which struggle to generalize to the intricacies of the Urdu script and the lack of sufficient annotated real-world data, we have introduced the UTRSet-Real, a large-scale annotated real-world dataset comprising over 11,000 lines and UTRSet-Synth, a synthetic dataset with 20,000 lines closely resembling real-world and made corrections to the ground truth of the existing IIITH dataset, making it a more reliable resource for future research. We also provide UrduDoc, a benchmark dataset for Urdu text line detection in scanned documents. Additionally, we have developed an online tool for end-to-end Urdu OCR from printed documents by integrating UTRNet with a text detection model. Our work not only addresses the current limitations of Urdu OCR but also paves the way for future research in this area and facilitates the continued advancement of Urdu OCR technology. The project page with source code, datasets, annotations, trained models, and online tool is available at abdur75648.github.io/UTRNet.

Keywords: Urdu OCR · UTRNet· UTRSet · Printed Text Recognition · High-Resolution Feature Extraction

1 Introduction

Printed text recognition, also known as optical character recognition (OCR), involves converting digital images of text into machine-readable text and is an important topic of research in the realm of document analysis with applications in a wide variety of areas [61]. While OCR has transformed the accessibility & utility

(a) Sample 1

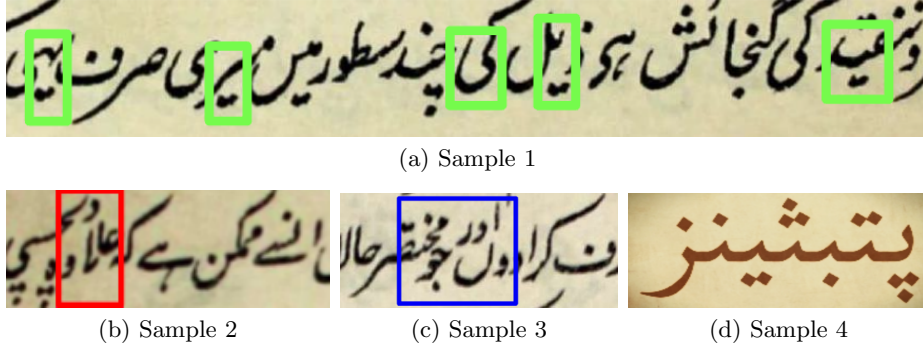(b) Sample 2   (c) Sample 3   (d) Sample 4

Fig. 1: Intricacies of the script. (a) All 5 characters inside the boxes are the same, but they look different as the shape of the character relies on its position and context in the ligature. The box contains 8 characters in (b) and 11 in (c), demonstrating that the script has a very high degree of overlap. (c) The word shown consists of 6 distinct characters of similar shape, which differ merely by the arrangement of little dots (called *"Nuqta"*) around them.

of written/printed information, it has traditionally been focused on Latin languages, leaving non-Latin low-resource languages such as Urdu, Arabic, Pashto, Sindhi and Persian largely untapped. Despite recent developments in Arabic script OCR [18,5,28], research on OCR for Urdu remains limited [34,27,35]. With over 230 million native speakers and a huge literature corpus, including classical prose and poetry, newspapers, and manuscripts, Urdu is the 10th most spoken language in the world [43,30]. Hence the development of a robust OCR system for Urdu remains an open research problem and a crucial requirement for efficient storage, indexing, and consumption of its vast heritage, mainly its classical literature.

However, the intricacies of the Urdu script, which predominantly exists in the Nastaleeq style, present significant challenges. It is primarily cursive, with a wide range of variations in writing style and a high degree of overall complexity, as shown in Fig. 1. Though Arabic script is similar [56], the challenges faced in recognizing Urdu text differ significantly. Arabic text is usually printed in the Naskh style, which is mostly upright and less cursive, and has only 28 alphabets [28], in contrast to the Urdu script, which consists of 45 main alphabets, 26 punctuation marks, 8 honorific marks, and 10 Urdu digits, as well as various special characters from Persian and Arabic, English alphabets, numerals, and punctuation marks, resulting in a total of 181 distinct glyphs [30] that need to be recognized. Furthermore, the lack of large annotated real-world datasets in Urdu compounds these challenges, making it difficult to compare different models' performance accurately and to continue advancing research in the field (Section 4). The lack of standardization in many Urdu fonts and their rendering schemes (particularly in early Urdu literature) further complicates the generation
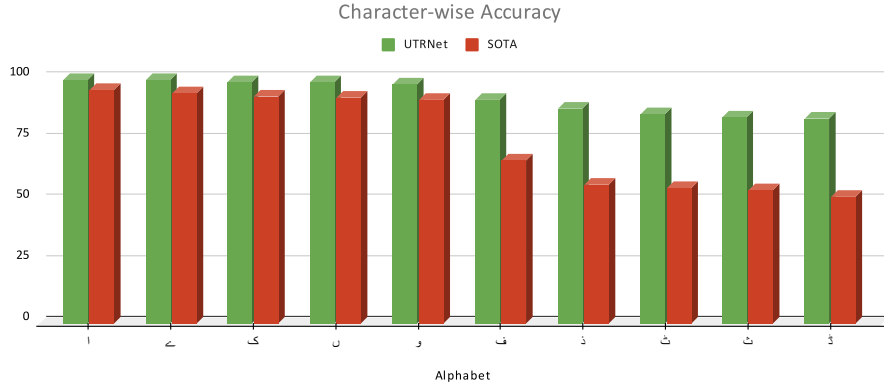
Character-wise Accuracy

Fig. 2: Plot showing character-wise accuracies of UTRNet-Small and SOTA for Urdu OCR, presented in [30]. It can be observed that the accuracy gap is larger for the last 5 characters (right side), which differ from several other characters having the same shape only in terms of presence and arrangement of dots around them, as compared to the first 5 characters, which are simpler ones.

of synthetic data that closely resembles real-world representations. This hinders experiments with more recent transformer-based OCR models that require large training datasets (Table 3B) [57,70,37]. As a result, a naive application of the methods developed for other languages does not result in a satisfactory performance for Urdu (Table 3), highlighting the need for exclusive research in OCR for Urdu.

The purpose of our research is to address these long-standing limitations in printed Urdu text recognition through the following key contributions:

1. We propose a novel approach using high-resolution, multi-scale semantic feature extraction in our UTRNet architecture, a hybrid CNN-RNN model, that demonstrates state-of-the-art performance on benchmark datasets.
2. We create UTRSet-Real, a large-scale annotated real-world dataset comprising over 11,000 lines.
3. We have developed a robust synthetic data generation module and release UTRSet-Synth, a high-quality synthetic dataset of 20,000 lines closely resembling real-world representations of Urdu text.
4. We correct many annotation errors in one of the benchmark datasets [30] for Urdu OCR, thereby elevating its reliability as a valuable resource for future research endeavours and release the corrected annotations publicly.
5. We have curated UrduDoc, a real-world Urdu documents text line detection dataset. The dataset is a byproduct of our efforts towards UTRSet-Real, and contains line segmentation annotation for 478 pages generated from more than 130 books.
6. To make the output of our project available to a larger non-computing research community, as well as lay users, we have developed an online tool for
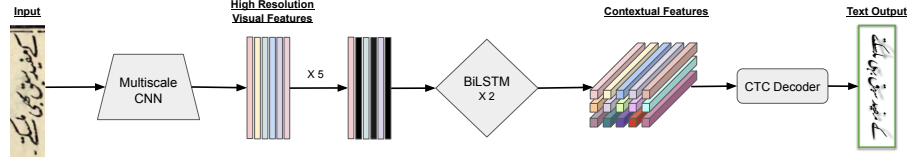
end-to-end Urdu OCR, integrating UTRNet with a third-party text detection model.

In addition to our key contributions outlined above, we conduct a thorough comparative analysis of state-of-the-art (SOTA) Urdu OCR models under similar experimental setups, using a unifying framework introduced by [8] that encompasses feature extraction, sequence modelling and prediction stages and provides a common perspective for all the existing methods. We also examine the contributions of individual modules towards accuracy as an ablation study (Table 2). Finally, we discuss the remaining limitations and potential avenues for future research in Urdu text recognition.
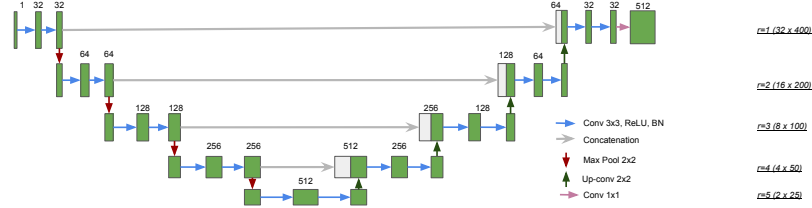
2   Related Work

The study of Urdu OCR has gained attention only in recent years. While the first OCR tools were developed more than five decades back [34], the earliest machine learning based Urdu OCR was developed in 2003 [47]. Since then, the research in this field has progressed from isolated character recognition to word/ligature level recognition to line level recognition (see [34,27,35] for a detailed survey). Early approaches primarily relied on handcrafted features, and traditional machine learning techniques, such as nearest neighbour classification, PCA & HMMs [64,33,53,2], to classify characters after first segmenting individual characters/glyphs from a line image. These techniques often required extensive pre-processing and struggled to achieve satisfactory performance on large, varied datasets.
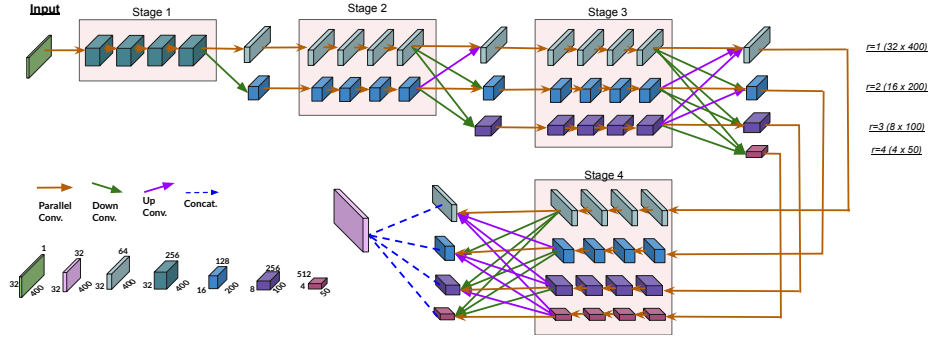
Recently segmentation-free end-to-end approaches based on CNN-RNN hybrid networks [30] have been introduced, in which a CNN [38] is used to extract low-level visual features from input data which is then fed to an RNN [63] to get contextual features for the output transcription layer. Among the current SOTA models for Urdu OCR (Table 3C), VGG networks [60] have been used for feature extraction in [30,45,44], whereas ResNet networks [25] have been utilized in [31]. For sequential modelling, BiLSTM networks [54] have been used in [30,31], while MDLSTM networks [13] have been employed in [45,44]. All of these approaches utilize a Connectionist Temporal Classification (CTC) layer [22] for output transcription. In contrast, [6] utilizes a DenseNet [26] and GRU network [16] with an attention-based decoder layer [9] for final transcription. Arabic, like Urdu, shares many similarities in the script as discussed above, and as such, the journey of OCR development has been similar [29,18]. Recent works have shown promising results in recognising Arabic text through a variety of methods, including traditional approaches [55,77], as well as DL-based approaches such as CNN-RNN hybrids [20,29], attention-based models [62,12], and a range of network architectures [49,23,36,32]. However, these approaches still struggle when it comes to recognizing Urdu text, as evident from the low accuracies achieved by SOTA Arabic OCR methods like [20], [62] and [12] in our experimental results presented in Table 3.

(a) Proposed overall architecture

(b) Multiscale feature extraction module for UTRNet-Small based on UNet [50]

(c) Multiscale feature extraction module for UTRNet-Large based on HRNet [66]

Fig. 3: Proposed Architecture

While each of the methods proposed so far has claimed to have pushed the boundary of the technology, they often rely on the same methodologies used for other languages without considering the complexities specific to Urdu script and, as such, do not fully utilize the potential in this field. Additionally, a fair comparison among these approaches has been largely missing due to inconsistencies in the datasets used, as described in the dataset section below.

3   Proposed Architecture

Recently, transformer-based models have achieved state-of-the-art performance on various benchmarks [24]. However, these models have the drawback of being highly data-intensive and requiring large amounts of training data, making it difficult to use them for several tasks with limited real-world data, such as

printed Urdu text recognition (as discussed in Section 1). In light of this, we propose UTRNet (Figure 3), a novel CNN-RNN hybrid network that offers a viable alternative. The proposed model effectively extracts high-resolution multiscale features while capturing sequential dependencies, making it an ideal fit for Urdu text recognition in real-world scenarios. We have designed UTRNet in two versions: UTRNet-Small (10.7M parameters) and UTRNet-Large (47.3M parameters). The architecture consists of three main stages: feature extraction, sequential modelling, and decoding. The feature extraction stage makes use of convolutional layers to extract feature representations from the input image. These representations are then passed to the sequential modelling stage to learn sequential dependencies among them. Finally, the decoding module converts the sequential data feature thus obtained into the final output sequence.

3.1   High-Resolution Multiscale Feature Extraction

In our proposed method, we address the wide mismatch in the accuracy of different characters observed in existing techniques, as shown in Figure 2. We posit that this is due to the lack of attention given to small features associated with most Urdu characters by the existing methods. These methods rely upon using standard CNNs, such as RCNN[21], VGG[60], ResNet[25], etc., as their backbone. However, a significant drawback of using these CNNs is the low resolution representation generated at the output layer. The representation lacks low-level feature information, such as the dots (called *"Nuqta"*) in Urdu characters. To overcome this limitation, in our proposed method, we propose to use a high-resolution multiscale feature extraction technique to extract features from an input image while preserving the small details of the image.

UTRNet-Small: To address the issue of computational efficiency, we propose a lighter variant of our novel UTRNet architecture, referred to as UTRNet-Small, which employs a standard U-Net model (as shown in Figure 3b), initially proposed in [50] for biomedical image segmentation. The lighter version proposed by us addresses captures high resolution feature maps using the standard U-Net model, originally proposed in [50] for biomedical image segmentation. We first encode the low-resolution representation of input image X, which captured context from a large receptive field. We then recovers the high-resolution representation using learnable convolutions from the previous decoder layer, and skip connections from the encoder layers. For any resolution with index $r \in \{1, 2, 3, 4, 5\}$, the feature map at that resolution can be defined as: $F_r = \text{downsample}(F_{r-1})$ during downsampling. Here downsample(F_{r-1}) is the downsampled feature map from the resolution index $(r-1)$. Similarly, $F_r = \text{concat}(\text{upsample}(F_{r+1}), M_r)$ represents feature map during upsampling, where upsample(F_{r+1}) is the upsampled feature map from the resolution index $(r+1)$, and M_r is the feature map from the downsampling path corresponding to that resolution). This allows the model to aggregate image features from multiple image scales.

UTRNet-Large: The model (Figure 3c) maintains high-resolution representation throughout the process and captures the fine-grained details more efficiently,

using an HRNet architecture [66]. The resulting network consists of 4 stages, with the Ith stage containing streams coming from $\#I$ different resolutions and giving out $\#I$ streams corresponding to the different resolutions. Each stream in the Ith stage is represented as R_r^I, where $r \in \{1, 2, \ldots, I\}$. These streams which are then inter-fused among themselves to get $\#(I + 1)$ final output streams, $R_r^{I,O}$ for the next stage:

$$R_r^{I,O} = \sum_{i=1}^{I} f_{ir}(R_i^I), \qquad r \in \{1, 2, \ldots, I + 1\}$$

The transform function f_{ir} is dependent on the input resolution index i and the output resolution index r. If $x = r$, then $f_{xr}(R) = R$. However, if $x < r$, then $f_{xr}(R)$ downsamples the input representation R. Similarly, if $x > r$, then $f_{xr}(R)$ upsamples the input resolution. UTRNet-Large uses repeated multi-resolution fusions which allows effective exchange of information across multiple resolutions. Thus, giving a multi-dimensional and high-resolution feature representation of the input image, which is semantically richer.

3.2   Sequential Modeling And Prediction

The output from the feature extraction stage is a feature map $V = \{v_i\}$. To prevent over-fitting, we implement a technique called Temporal Dropout [14], in which we randomly drop half of the visual features before passing them to the next stage. We do this 5 times in parallel and take the average. In order to capture the rich contextual information and temporal relationships between the features thus obtained, we pass it through 2 layers of BiLSTM [54] (DBiLSTM) [58]. Each BiLSTM layer identifies two hidden states, h_t^f and h_t^b, calculated forward and backward through time, respectively, which are combined to determine one hidden state h^t, using an FC layer. The sequence $H = \{h_t\} = \text{DBiLSTM}(V))$ thus obtained has rich contextual information from both directions, which is crucial for Urdu text recognition, especially because the shape of each character depends upon characters around it. The final prediction output $Y = \{y1, y2, \ldots\}$, a variable-length sequence of characters, is generated by the prediction module from the input sequence H. For this, we use the Connectionist temporal classification (CTC), as described in [22].

4   Current Publicly Available and Proposed Datasets

The availability of datasets for the study of Urdu Optical Character Recognition (OCR) is limited, with only a total of six datasets currently available: UPTI [51], IIITH [30], UNHD [1], CENPARMI [52], CALAM [17], and PMU-UD [3]. Of these datasets, only IIITH and UPTI contain printed text line samples, out of which only UPTI has a sufficient number of samples for training. However, the UPTI dataset is synthetic in nature, with limited diversity and simplicity in

(a) IIITH [30]   (b) UPTI [51]

(c) Proposed UTRSet-Synth

(d) Proposed UTRSet-Real

Dataset	Training Set	Validation Set	Vocab Length	Type
IIITH [30]	NA	1,610	5,772	Real
UPTI [51]	8,051	2,012	12,054	Synthetic
UTRSet-Real	9,065	2,096	22,964	Real
UTRSet-Synth	20,000	NA	28,187	Synthetic

Fig. 4: Sample images, and statistics of the publicly available and proposed datasets. Notice the richness and diversity in the proposed UTRSet-Real and UTRSet-Synth datasets, as compared to existing datasets

comparison to real-world images (See Figure 4). There is currently no comprehensive real-world printed Urdu OCR dataset available publicly for researchers. As a result, different studies have used their own proprietary datasets, making it difficult to determine the extent to which proposed models improve upon existing approaches. This lack of standardisation and transparency hinders the ability to compare the performance of different models accurately and to continue advancing research in the field, which our work aims to tackle with the introduction of two new datasets. The following are the datasets used in this paper (also summarized in Figure 4):

IIITH: This dataset was introduced by [30] in 2017, and it contains only the validation data set of 1610 line images in nearly uniform colour and font. No training data set has been provided. We corrected the ground-truth annotations for this dataset, as we found several mistakes, as highlighted in Figure 5.

UPTI: Unlike the other two, this is a synthetic data set introduced in 2013 by [51]. It consists of a total of 10,063 samples, out of which 2,012 samples are in
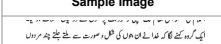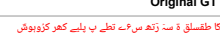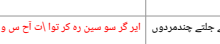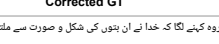
Sample Image	Original GT	Corrected GT
ایک ہو کے کاگ نی یا مان کی گل، صورت سے شکل چلا پڑ مردوں ع، ایٹ ع بس بن بن ربا رر بی، سوٹ رت چپ بنی جلتے	ایر گر سو سین رہ کر توا اُت آح س و کا طفسلق ة مٰ زٹھ س۶ے تعلی پ پلیے کھر کزوبوش	ایک گروہ کینی لگا کہ خدا نے ان بتوں کی شکل و صورت سے ملتے جلتے چندمردوں
ساتھ کنے لگے "ماراینا مثل حجتک یا محمد شهداللہ رسول اللہ"اے محمد	:ساتھ کینی لگے : عمادائنا مثلی حجضت یا مضل شهلماتندصول اللهﷺا اے مخ	ساتھ کینی لگے: "ماراینا مثل حجتک یا محمد شهداللہ رسول اللہ"اے محمد
ہجرت مدینہ سے پہلے یہودی آبس میں ان نشانیوں کا تذکرہ کیا کارتے	ه هیندا پھ .علر ۔بل، و ٮن او روۇ ھا اگں ٮر ۓ۔	ہجرت مدینہ سے پہلی یہودی آبس میں ان نشانیوں کا تزکرہ کیا کارتی
الناس في أحكام دینهم في شرق الأرض وغربها، في هٰذا	:الناس فذب اهکاء دینا ، فبا شرفی ادٰ وغر بها"، فٮچ حذه	,الناس فی احکام دینهم فی شرق الارض ع غیربها.. فی هذه العصور
۲۰۰۲/۳/۲٤م رقم ۱۸۰٤٦..	دقم ٣٩	م رقم ۱۸۰۴۶/ـ۲/۲/۲۰۰۲/۲۴
رِّبَك بالحِكْمَـٰة والْمَوْعِظَـٰة الْحَسَنَـٰة وَجَـٰدِلْهُمْ بِـٰأَتِي هَـٰيَ	دٮگل بالٮٰ: واکو تعفظا' الصهٮ، و کجاده اہ : با اتی ٮیہٰ:	ربك بالحكمة والفوعظة الخسنة وَجادلهم بالتی هی

Fig. 5: Mistakes in IIITH dataset annotations and our corrections.

the validation set. This data set is also uniform in colour and font and has a vocabulary of 12,054 words.

Proposed UTRSet-Real: A comprehensive real-world annotated dataset curated by us, containing a few easy and mostly hard images (as illustrated in Figure 1). To create this dataset, we collected 130 books and old documents, scanned over 500 pages, and manually annotated the scanned images with linewise bounding boxes and corresponding ground truth labels. After cropping the lines and performing data cleaning, we obtained a final dataset of 11,161 lines, with 2,096 lines in the validation set and the remaining in the training set. This dataset stands out for its diversity, with various fonts, text sizes, colours, orientations, lighting conditions, noises, styles, and backgrounds represented in the samples, making it well-suited for real-world Urdu text recognition.

Proposed UTRSet-Synth: To complement the real-world data in UTRSet-Real for training purposes, we also present UTRSet-Synth, a high-quality synthetic dataset of 20,000 lines with over 28,000 total unique words, closely resembling real-world representations of Urdu text. This dataset is generated using a custom-designed synthetic data generation module that allows for precise control over variations in font, text size, colour, resolution, orientation, noise, style, background etc. The module addresses the challenge of standardizing fonts by collecting and incorporating over 130 diverse fonts of Urdu after making corrections to their rendering schemes. It also addresses the limitation of current datasets, which have very few instances of Arabic words and numerals, Urdu digits etc., by incorporating such samples in sufficient numbers. Additionally, it generates text samples by randomly selecting words from a vocabulary of 100,000 words. The generated UTRSet-Synth has 28,187 unique words with an average word length of 7 characters. The data generation module has been made publicly available on the project page to facilitate further research.

Proposed UrduDoc: In addition to the recognition datasets discussed above, we also present UrduDoc, a benchmark dataset for Urdu text line detection in scanned documents. To the best of our knowledge, this is the first dataset of its kind [4,15]. It was created as a byproduct of the UTRSet-Real dataset generation process, in which the pages were initially scanned and then annotated with horizontal bounding boxes in COCO format [41] to crop the text lines. Comprising of 478 diverse images collected from various sources such as books, documents,

Fig. 6: Sample images from the UrduDoc Dataset: An annotated real-world benchmark for Urdu text line detection in scanned documents

manuscripts, and newspapers, it is split into 358 pages for training and 120 for validation. The images include a wide range of styles, scales, and lighting conditions, making them a valuable resource for the research community. The UrduDoc dataset will serve as a valuable resource for the research community, advancing research in Urdu document analysis. We also provide benchmark results of a few SOTA text detection models on this dataset using precision, recall, and h-mean, as shown in Table 1. The results demonstrate that the Contour-Net model [69] outperforms the other models in terms of h-mean. It is worth noting that as text detection was not the primary focus of our research but rather a secondary contribution, a thorough examination of text detection has not been conducted. This aspect can be considered a future work for researchers interested in advancing the field of Urdu document analysis. We will make this dataset publicly available to the research community for non-commercial, academic, and research purposes associated with Urdu document analysis, subject to request and upon execution of a no-cost license agreement.

5   Experiments And Results

Experimental Setup: In order to ensure a fair comparison among the existing models in the field, we have established a consistent experimental setup for evaluating the performance of all the models and report all the results in Table

Methods	Precision	Recall	Hmean
EAST [75]	70.43	72.56	71.48
PSENet [67]	78.32	77.91	78.11
DRRG [72]	83.05	84.72	83.87
ContourNet [69]	85.36	88.68	86.99

Table 1: Experimental results on UrduDoc

Model/Strategy	UTRSet-Real	IIITH	UPTI
Various Multiscale Feature Extraction Backbones			
UNet [50]	90.87	86.35	95.08
AttnNet [46]	91.95	86.45	95.26
ResidualUNet [73]	91.90	87.16	95.23
InceptionUNet [48]	92.10	87.37	95.61
UNetPlusPlus [76]	92.53	87.36	95.84
HRNet [66]	**92.97**	**88.01**	**95.97**
Various Sequential Modelling Backbones			
LSTM	91.20	87.21	94.41
GRU	91.48	87.04	94.58
MDLSTM	91.51	87.38	94.67
BiLSTM	91.53	87.67	94.69
DBiLSTM	**92.97**	**88.01**	**95.97**
Various Strategies to Improve Generalization			
None	91.07	86.80	95.07
Augmentation	92.35	87.53	95.48
Augmentation + Temporal Dropout	**92.97**	**88.01**	**95.97**

Table 2: The results of the ablation study for UTRNet provide a comprehensive examination of the impact of individual feature extraction and sequence modelling stages and augmentation strategies on the overall performance of the model. By evaluating each stage one-by-one while keeping the remaining stages constant, the results highlight the key factors driving the accuracy of the model, offering valuable insight into optimizing the performance of UTRNet.

3. Specifically, we have fixed the choice of training to the UTRSet-Real training set, the validation set to be the validation sets of the datasets outlined in Figure 4. Further, to compare the different available training datasets, we train our proposed UTRNet models on each of them and present the results in Table 4. We have used the AdaDelta optimizer [71], with a decay rate of 0.95, to train our UTRNet models on an NVIDIA A100 40GB GPU. The batch size and learning rate used were 32 and 1.0, respectively. Gradient clipping was employed at a magnitude of 5, and all parameters were initialized using He's method [66]. To improve the robustness of the model, we employed a variety of data augmentation techniques during the training process, such as random resizing, stretching/compressing, rotation, and translation, various types of noise, random border crop, contrast stretching, and various image processing techniques, to simulate different types of imaging conditions and improve the model's ability to generalize to real-world scenarios. The UTRNet-Large model achieved convergence in approximately 7 hours. We utilize the standard character-wise accuracy metric for comparison, which uses the edit distance between the predicted output

Models	UTRSet-Real	IIITH	UPTI
A. Baseline OCR models (Hybrid CNN-RNN)			
R2AM [39]	84.12	81.39	92.07
CRNN [58]	83.11	81.45	91.49
GRCNN [65]	84.21	81.09	92.28
Rosetta [11]	84.08	81.94	92.15
RARE [59]	85.63	83.59	92.74
STAR-Net [42]	87.05	84.27	93.59
TRBA [8]	88.92	85.61	94.16
B. Baseline OCR models (Transformer-based)			
Parseq [10]	26.13	25.60	26.41
ViTSTR [7]	34.86	32.63	35.78
TrOCR [40]	38.43	36.10	37.61
ABINet [19]	41.17	40.20	38.96
CDistNet [74]	33.72	34.96	32.48
VisionLAN [68]	28.40	27.82	29.07
C. SOTA Urdu OCR models			
5LayerCNN-DBiLSTM [20]	82.92	81.15	90.67
VGG-BiLSTM [30]	83.11	81.45	91.49
VGG-MDLSTM [45,44]	83.30	81.72	91.17
VGG-LSTM-Attn [62]	84.16	82.21	91.88
VGG-DBiLSTM-Attn [12]	84.58	82.72	92.01
ResNet-BiLSTM [31]	86.96	84.18	93.61
DenseNet-GRU-Attn [6]	91.10	85.32	94.63
UTRNet-Small	90.87	86.35	95.08
UTRNet	**92.97**	**88.01**	**95.97**

Table 3: Performance comparison of SOTA OCR models

(Pred) and the ground truth (GT):

$$\text{Accuracy} = \frac{\sum \left(\text{length}(GT) - \text{EditDistance}(Pred, GT)\right)}{\sum \left(\text{length}(GT)\right)}$$

5.1   Results And Analysis

In order to evaluate the effectiveness of our proposed architecture, we conducted a series of experiments and compared our results with state-of-the-art (SOTA) models for Urdu OCR, as well as a few for Arabic, including both printed and handwritten ones (Table 3C). Additionally, we evaluated our model against the current SOTA baseline OCR models, primarily developed for Latin-based languages (Tables 3A and 3B). Our proposed model achieves superior performance, surpassing all of the SOTA OCR models in terms of character-wise accuracy on all three validation datasets, achieving a recognition accuracy of 92.97% on the UTRSet-Real validation set. It is worth noting that while Table 3A presents

a comparison of our proposed model against hybrid CNN-RNN models, Table 3B presents a comparison against recent transformer-based models. The results clearly show that transformer-based models perform poorly in comparison to our proposed model and even the SOTA CNN-RNN models for Latin OCR. This can be attributed to the fact that these models, which are designed to be trained on massive datasets when applied to the case of Urdu script recognition, overfit the small-size training data and struggle to generalize, thus resulting in poor validation accuracy.

In our analysis of our proposed UTRSet-Real and UTRSet-Synth datasets against the existing UPTI dataset, which is currently the only available training dataset for this purpose, we found that our proposed datasets effectively improve the performance of the UTRNet model. When trained on the UPTI dataset, both UTRNet-Small and UTRNet achieve high accuracy on the UPTI validation set but perform poorly on the UTRSet-Real and IIITH validation sets. This suggests that the UPTI dataset is not representative of real-world scenarios and does not adequately capture the complexity and diversity of printed Urdu text. Our proposed datasets, on the other hand, are specifically designed to address these issues and provide a more comprehensive and realistic representation of the task at hand, as both of them perform significantly well on all datasets, with UTRSet-Real being the best. Furthermore, the results show that combining all three training datasets can further improve the performance, especially on the IIITH and UPTI validation sets.

One of the key insights from our results is the significant difference in accuracy when comparing our proposed UTRNet model with the current state-of-the-art (SOTA) model for printed Urdu OCR [30], as presented in Figure 2. This highlights the complexity of recognizing the intricate features of Urdu characters and the efficacy of our proposed UTRNet model in addressing these challenges. Our results align with our hypothesis that high-resolution multi-scale feature maps are essential for capturing the nuanced details required for accurate Urdu OCR. To further support this claim, we also present a visualization of feature maps generated from our CNN (as depicted in Figure 7), which

Model	Training Data	UTRSet-Real	IIITH	UPTI
UTRNet-Small	UPTI	54.84	72.82	98.63
UTRNet-Small	UTRSet-Real	90.87	86.35	95.08
UTRNet-Small	UTRSet-Synth	75.14	85.47	92.09
UTRNet-Small	Mix-All	91.71	90.04	98.72
UTRNet	UPTI	64.73	76.15	**99.17**
UTRNet	UTRSet-Real	92.97	88.01	95.97
UTRNet	UTRSet-Synth	80.26	87.85	93.49
UTRNet	Mix-All	**93.39**	**90.91**	98.36

Table 4: Performance after training with different datasets. See that training with UPTI dataset leads to poor performance on real-world validation datasets. (Here, "Mix-All" means a mixture of UPTI, UTRSet-Real & UTRSet-Synth)
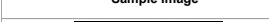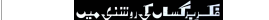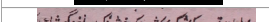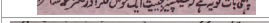
Fig. 7: Visualization of feature maps learnt by various layers of the UTRNet-Large. The small features associated with various characters are lost at low resolution (middle row) but are present in the last row. This illustrates the effectiveness of our proposed high-resolution multi-scale feature extraction technique, leading to improved performance in the printed Urdu text recognition.

clearly demonstrates the ability of UTRNet to effectively extract and preserve the high-resolution details of the input image. Additionally, we provide a qualitative analysis of the results by comparing our model with the SOTA [30] in Figure 9.

We also conducted a series of ablation studies to investigate the impact of various components of our proposed UTRNet model on performance. The results of this study, as shown in Table 2, indicate that each component of our model makes a significant contribution to the overall performance. Specifically, since we observed that incorporating a multi-scale high-resolution feature extraction significantly improves the result for Urdu OCR, we tried various other multi-scale CNNs as a part of our ablation study. We also found that the use of generalisation techniques, such as temporal dropout and augmentation, further improved the robustness of our model, making it able to effectively handle a wide range of challenges which are commonly encountered in real-world Urdu OCR scenarios.

6 Web Tool for End-to-End Urdu OCR

We have developed an online website for Urdu OCR that integrates ContourNet [69] model, trained on the Ur-

Fig. 8: Web tool developed by us for end-to-end Urdu OCR

Sample Image	SOTA Prediction	UTRNet Prediction
قلسترپرگساب کی روشنئی میب	عاریرکساپ کی رونئنی میپ	قل بر گساں کی روننئی میں
پہلی بات قویبہ کرشیکسیپہ پختیت ایک خوش لگا اور نغز گوشاعر ا	پہلی بات تو یبے ک یکیت کیت ابک خوش ک اور فزگرشاع ا	پہلی بات تو ی۔ بے ک۔ شیکبیر بحثیت ایک خوش فکر اور نغز گوشاعر
حالت ہمارہ کاہی گدہ جوکچہ اسرا اغا نفطا فظا ۔ اس کی نۃ ئی بیوی تھی اور د نہ یجے۔	حالت باری تں ک بچ کم ھ ا ا کپ۔ بی ی ہ بھی ے	حالت بنار بی تھی ک۔ وہ جوکچھ ک۔ رباتھا غلط تھا۔ اس کی ن۔ تو بیوی تھی اور ن۔ چے۔
سے دلداد گی کا رجحن او ر شاعرانہ رندی دبنا گی سے دور رئ کا اظہارا	ا دلاداری پگ ای ا ا ای ی ی وا ین ک یو ای م ل لم	سے دلدارگی کا رجحان ادرشاعرانہ رندی دبیا کی سے دوری کا اظہار
تکھرتھیم دبا بے ۔ 1	کلپ۔ ثرھ د با بے:1	"اکلم۔ پڑھ ربا بی۔

Fig. 9: The figure illustrates a qualitative analysis of our proposed UTRNet-Small and the SOTA method [30] on the UTRSet-Real dataset. To facilitate a fair comparison, the errors in the transcriptions are highlighted in red. The results showcase the superior accuracy of UTRNet in capturing fine-grained details and accurately transcribing Urdu text in real-world documents.

duDoc dataset with our proposed UTRNet model. This integration allows for end-to-end text recognition in real-world documents, making the website a valuable tool for easy and efficient digitization of a large corpus of available Urdu literature.

7 Conclusion

In this work, we have addressed the limitations of previous works in Urdu OCR, which struggle to generalize to the intricacies of the Urdu script and the lack of large annotated real-world data. We have presented a novel approach through the introduction of a high-resolution, multi-scale semantic feature extraction-based model which outperforms previous SOTA models for Urdu OCR, as well as Latin OCR, by a significant margin. We have also introduced three comprehensive datasets: UTRSet-Real, UTRSet-Synth, and UrduDoc, which are significant contributions towards advancing research in printed Urdu text recognition. Additionally, the corrections made to the ground truth of the existing IIITH dataset have made it a more reliable resource for future research. Furthermore, we've also developed a web based tool for end-to-end Urdu OCR which we hope will help in digitizing the large corpus of available Urdu literature. Despite the promising results of our proposed approach, there remains scope for further optimization and advancements in the field of Urdu OCR. A crucial area of focus is harnessing the power of transformer-based models along with large amounts of synthetic data by enhancing the robustness and realism of synthetic data and potentially achieving even greater performance gains. Our work has laid the foundation for continued progress in this field, and we hope it will inspire new and innovative approaches for printed Urdu text recognition.

8   Acknowledgement

References

1. Ahmed, S.B., Naz, S., Swati, S., Razzak, M.I.: Handwritten urdu character recognition using 1-dimensional blstm classifier (2017). https://doi.org/10.48550/ARXIV.1705.05455
2. Akram, M.U., Hussain, S.: Word segmentation for urdu ocr system (2010)
3. Alghazo, J.M., Latif, G., Alzubaidi, L., Elhassan, A.: Multi-language handwritten digits recognition based on novel structural features. Journal of Imaging Science and Technology (2019)
4. Ali, A., Pickering, M.: Urdu-text: A dataset and benchmark for urdu text detection and recognition in natural scenes. 2019 International Conference on Document Analysis and Recognition (ICDAR) pp. 323–328 (2019). https://doi.org/10.1109/ICDAR.2019.00059
5. Althobaiti, H., Lu, C.: A survey on arabic optical character recognition and an isolated handwritten arabic character recognition algorithm using encoded freeman chain code. In: 2017 51st Annual Conference on Information Sciences and Systems (CISS). pp. 1–6 (2017). https://doi.org/10.1109/CISS.2017.7926062
6. Anjum, T., Khan, N.: An attention based method for offline handwritten urdu text recognition. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 169–174 (2020). https://doi.org/10.1109/ICFHR2020.2020.00040
7. Atienza, R.: Vision transformer for fast and efficient scene text recognition (2021). https://doi.org/10.48550/ARXIV.2105.08582, https://arxiv.org/abs/2105.08582
8. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis (2019). https://doi.org/10.48550/ARXIV.1904.01906
9. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014). https://doi.org/10.48550/ARXIV.1409.0473
10. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models (2022). https://doi.org/10.48550/ARXIV.2207.06966, https://arxiv.org/abs/2207.06966
11. Borisyuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM (2018). https://doi.org/10.1145/3219819.3219861, https://doi.org/10.11452F3219819.3219861
12. Butt, H., Raza, M.R., Ramzan, M., Ali, M.J., Haris, M.: Attention-based cnn-rnn arabic text recognition from natural scene images. Forecasting **3**, 520–540 (07 2021). https://doi.org/10.3390/forecast3030033

13. Byeon, W., Liwicki, M., Breuel, T.M.: Texture classification using 2d lstm networks. In: 2014 22nd International Conference on Pattern Recognition. pp. 1144–1149 (2014). https://doi.org/10.1109/ICPR.2014.206

14. Chammas, E., Mokbel, C.: Fine-tuning handwriting recognition systems with temporal dropout. ArXiv **abs/2102.00511** (2021)

15. Chandio, A.A., Asikuzzaman, M., Pickering, M., Leghari, M.: Cursive-text: A comprehensive dataset for end-to-end urdu text recognition in natural scene images. Data in Brief **31**, 105749 (2020). https://doi.org/https://doi.org/10.1016/j.dib.2020.105749

16. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches (2014). https://doi.org/10.48550/ARXIV.1409.1259

17. Choudhary, P., Nain, N.: A four-tier annotated urdu handwritten text image dataset for multidisciplinary research on urdu script. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **15**(4) (may 2016). https://doi.org/10.1145/2857053

18. Djaghbellou, S., Bouziane, A., Attia, A., Akhtar, Z.: A survey on arabic handwritten script recognition systems. International Journal of Artificial Intelligence and Machine Learning **11**, 1–17 (07 2021). https://doi.org/10.4018/IJAIML.20210701.oa9

19. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition (2021). https://doi.org/10.48550/ARXIV.2103.06495, https://arxiv.org/abs/2103.06495

20. Fasha, M., Hammo, B.H., Obeid, N., Widian, J.: A hybrid deep learning model for arabic text recognition. ArXiv **abs/2009.01987** (2020)

21. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation (2013). https://doi.org/10.48550/ARXIV.1311.2524

22. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning. p. 369–376. ICML '06 (2006). https://doi.org/10.1145/1143844.1143891

23. Graves, A., Schmidhuber, J.: Offline arabic handwriting recognition with multidimensional recurrent neural networks. pp. 545–552 (01 2008)

24. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 87–110 (2023). https://doi.org/10.1109/TPAMI.2022.3152247

25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). https://doi.org/10.48550/ARXIV.1512.03385

26. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2016). https://doi.org/10.48550/ARXIV.1608.06993

27. Husnain, M., Saad Missen, M.M., Mumtaz, S., Coustaty, M., Luqman, M., Ogier, J.M.: Urdu handwritten text recognition: a survey. IET Image Processing **14**(11), 2291–2300 (2020). https://doi.org/https://doi.org/10.1049/iet-ipr.2019.0401

28. Hussain, S.: A survey of ocr in arabic language: Applications, techniques, and challenges. Applied Sciences **13**, 27 (04 2023). https://doi.org/10.3390/app13074584

29. Jain, M., Mathew, M., Jawahar, C.V.: Unconstrained scene text and video text recognition for arabic script. 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR) pp. 26–30 (2017)

30. Jain, M., Mathew, M., Jawahar, C.: Unconstrained ocr for urdu using deep cnn-rnn hybrid networks. In: 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR). pp. 747–752. IEEE (2017)
31. Kashif, M.: Urdu handwritten text recognition using resnet18 (2021). https://doi.org/10.48550/ARXIV.2103.05105
32. Kassem, A.M., Mohamed, O., Ashraf, A., Elbehery, A., Jamal, S., Khoriba, G., Ghoneim, A.S.: Ocformer: A transformer-based model for arabic handwritten text recognition. 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) pp. 182–186 (2021)
33. Khan, K., Ullah, R., Ahmad, N., Naveed, K.: Urdu character recognition using principal component analysis. International Journal of Computer Applications **60** (12 2012). https://doi.org/10.5120/9733-2082
34. Khan, N.H., Adnan, A.: Urdu optical character recognition systems: Present contributions and future directions. IEEE Access **6**, 46019–46046 (2018). https://doi.org/10.1109/ACCESS.2018.2865532
35. Khan, N.H., Adnan, A., Basar, S.: An analysis of off-line and on-line approaches in urdu character recognition. In: 2016 15th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'16) (2016)
36. Ko, D., Lee, C., Han, D., Ohk, H., Kang, K., Han, S.: Approach for machine-printed arabic character recognition: the-state-of-the-art deep-learning method. electronic imaging **2018**, 176–1–176–8 (2018)
37. Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., Zhai, X.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
38. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
39. Lee, C.Y., Osindero, S.: Recursive recurrent nets with attention modeling for ocr in the wild (2016). https://doi.org/10.48550/ARXIV.1603.03101, https://arxiv.org/abs/1603.03101
40. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models (2021). https://doi.org/10.48550/ARXIV.2109.10282, https://arxiv.org/abs/2109.10282
41. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2014). https://doi.org/10.48550/ARXIV.1405.0312
42. Liu, W., Chen, C., Wong, K.Y., Su, Z., Han, J.: Star-net: A spatial attention residue network for scene text recognition. pp. 43.1–43.13 (01 2016). https://doi.org/10.5244/C.30.43
43. Mushtaq, F., Misgar, M.M., Kumar, M., Khurana, S.S.: UrduDeepNet: offline handwritten urdu character recognition using deep neural network. Neural Computing and Applications **33**(22), 15229–15252 (Nov 2021)
44. Naz, S., Ahmed, S., Ahmad, R., Razzak, M.: Zoning features and 2dlstm for urdu text-line recognition. Procedia Computer Science **96**, 16–22 (09 2016). https://doi.org/10.1016/j.procs.2016.08.084
45. Naz, S., Umar, A.I., Ahmad, R., Siddiqi, I., Ahmed, S.B., Razzak, M.I., Shafait, F.: Urdu nastaliq recognition using convolutional–recursive deep learning. Neurocomputing **243**, 80–87 (2017). https://doi.org/https://doi.org/10.1016/j.neucom.2017.02.081, https://www.sciencedirect.com/science/article/pii/S0925231217304654

46. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas (2018). https://doi.org/10.48550/ARXIV.1804.03999

47. Pal, U., Sarkar, A.: Recognition of printed urdu script. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. pp. 1183–1187 (2003). https://doi.org/10.1109/ICDAR.2003.1227844

48. Punn, N.S., Agarwal, S.: Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images. ACM Trans. Multimedia Comput. Commun. Appl. **16**(1) (feb 2020). https://doi.org/10.1145/3376922

49. Rashid, S.F., Schambach, M.P., Rottland, J., Nüll, S.: Low resolution arabic recognition with multidimensional recurrent neural networks (08 2013). https://doi.org/10.1145/2505377.2505385

50. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)

51. Sabbour, N., Shafait, F.: A segmentation free approach to arabic and urdu ocr. Proceedings of SPIE - The International Society for Optical Engineering **8658** (02 2013). https://doi.org/10.1117/12.2003731

52. Sagheer, M.W., He, C.L., Nobile, N., Suen, C.Y.: A new large urdu database for off-line handwriting recognition. In: Foggia, P., Sansone, C., Vento, M. (eds.) Image Analysis and Processing – ICIAP 2009. pp. 538–546. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)

53. Sardar, S., Wahab, A.: Optical character recognition system for urdu. 2010 International Conference on Information and Emerging Technologies pp. 1–5 (2010)

54. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)

55. Semary, N., Rashad, M.: Isolated printed arabic character recognition using knn and random forest tree classifiers. vol. 488, pp. 11– (11 2014)

56. Shahin, A.: Printed arabic text recognition using linear and nonlinear regression. International Journal of Advanced Computer Science and Applications **8** (02 2017). https://doi.org/10.14569/IJACSA.2017.080129

57. Shaiq, M.D., Cheema, M.D.A., Kamal, A.: Transformer based urdu handwritten text optical character reader (2022). https://doi.org/10.48550/ARXIV.2206.04575, https://arxiv.org/abs/2206.04575

58. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition (2015). https://doi.org/10.48550/ARXIV.1507.05717

59. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification (2016). https://doi.org/10.48550/ARXIV.1603.03915, https://arxiv.org/abs/1603.03915

60. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). https://doi.org/10.48550/ARXIV.1409.1556

61. Singh, A., Bacchuwar, K., Bhasin, A.: A survey of ocr applications. International Journal of Machine Learning and Computing (IJMLC) (01 2012). https://doi.org/10.7763/IJMLC.2012.V2.137

62. Sobhi, M., Hifny, Y., Mesbah Elkaffas, S.: Arabic optical character recognition using attention based encoder-decoder architecture. In: 2020 2nd International Conference on Artificial Intelligence, Robotics and Control. p. 1–5. AIRC'20, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3448326.3448327, https://doi.org/10.1145/3448326.3448327

63. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks (2014). https://doi.org/10.48550/ARXIV.1409.3215, https://arxiv.org/abs/1409.3215
64. Tabassam, N., Naqvi, S., Rehman, H., Anoshia, F.: Optical character recognition system for urdu (naskh font) using pattern matching technique. International Journal of Image Processing **3** (09 2009)
65. Wang, J., Hu, X.: Gated recurrent convolution neural network for ocr. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 334–343. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
66. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition (2019)
67. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network (2019). https://doi.org/10.48550/ARXIV.1903.12473
68. Wang, Y., Xie, H., Fang, S., Wang, J., Zhu, S., Zhang, Y.: From two to one: A new scene text recognizer with visual language modeling network (2021). https://doi.org/10.48550/ARXIV.2108.09661, https://arxiv.org/abs/2108.09661
69. Wang, Y., Xie, H., Zha, Z., Xing, M., Fu, Z., Zhang, Y.: Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection (2020). https://doi.org/10.48550/ARXIV.2004.04940
70. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet (2021). https://doi.org/10.48550/ARXIV.2101.11986, https://arxiv.org/abs/2101.11986
71. Zeiler, M.D.: Adadelta: An adaptive learning rate method (2012). https://doi.org/10.48550/ARXIV.1212.5701
72. Zhang, S.X., Zhu, X., Hou, J.B., Liu, C., Yang, C., Wang, H., Yin, X.C.: Deep relational reasoning graph network for arbitrary shape text detection (2020). https://doi.org/10.48550/ARXIV.2003.07493
73. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (may 2018). https://doi.org/10.1109/lgrs.2018.2802944, https://doi.org/10.11092Flgrs.2018.2802944
74. Zheng, T., Chen, Z., Fang, S., Xie, H., Jiang, Y.G.: Cdistnet: Perceiving multi-domain character distance for robust text recognition (2021). https://doi.org/10.48550/ARXIV.2111.11011, https://arxiv.org/abs/2111.11011
75. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: An efficient and accurate scene text detector (2017). https://doi.org/10.48550/ARXIV.1704.03155
76. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation (2018). https://doi.org/10.48550/ARXIV.1807.10165
77. Zoizou, A., Zarghili, A., Chaker, I.: A new hybrid method for arabic multi-font text segmentation, and a reference corpus construction. J. King Saud Univ. Comput. Inf. Sci. **32**, 576–582 (2020)