IET Image Processing

Research Article

Recognition of cursive video text using a deep learning framework

ISSN 1751-9659
Received on 30th September 2019
Revised 12th March 2020
Accepted on 30th June 2020
E-First on 14th October 2020
doi: 10.1049/iet-ipr.2019.1070
www.ietdl.org

Ali Mirza[1] ✉, *Imran Siddiqi*[1]

[1] *Department of Computer Science, Bahria University, Islamabad, Pakistan*

✉ *E-mail: alimirza@bahria.edu.pk*

Abstract: This study focuses on cursive text recognition appearing in videos, using a complete framework of deep neural networks. While mature video optical character recognition systems (V-OCRs) are available for text in non-cursive scripts, recognition of cursive scripts is marked by many challenges. These include complex and overlapping ligatures, context-dependent shape variations and presence of a large number of dots and diacritics. The authors present an analytical technique for recognition of cursive caption text that relies on a combination of convolutional and recurrent neural networks trained in an end-to-end framework. Text lines extracted from video frames are preprocessed to segment the background and are fed to a convolutional neural network for feature extraction. The extracted feature sequences are fed to different variants of bi-directional recurrent neural networks along with the ground truth transcription to learn sequence-to-sequence mapping. Finally, a connectionist temporal classification layer is employed to produce the final transcription. Experiments on a data set of more than 40,000 text lines from 11,192 video frames of various News channel videos reported an overall character recognition rate of 97.63%. The proposed work employs Urdu text as a case study but the findings can be generalised to other cursive scripts as well.

1 Introduction

With the tremendous increase in the amount of digital video data, the traditional tag-based video search engines need to be replaced with intelligent content-based retrieval systems. The conventional video retrieval systems rely on matching the queried words with user-assigned annotations and, ignore the rich information in videos that can be exploited for effective indexing and subsequent retrieval. The content-based search systems may exploit the visual information (objects and buildings), audio (spoken words), textual content (news tickers, names, subtitles etc.) or a combination of these to support smart retrieval. Examples of typical queries to such intelligent systems include retrieving all videos where a particular individual has appeared or all instances where a particular keyword (for example Breaking News) has been flashed. Among various search modalities, the focus of our present study lies on the textual content appearing in videos.

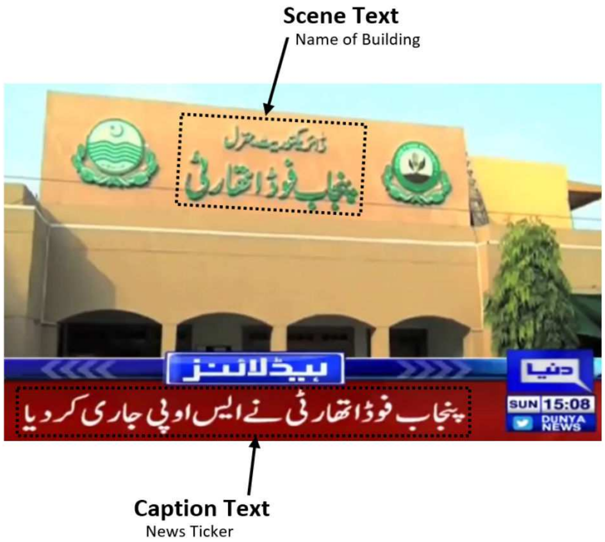


Fig. 1 *Video frame with instances of scene and caption text*

Textual content in videos can be categorised into two different classes, scene text and caption text (also known as artificial/graphics text). Scene text is captured through the camera during video recording and may not always be co-related with the content. Examples of scene text include signboards, advertisement banners, building names, text on T-shirts and play cards etc. It may also include handwritten text. Scene text is useful for applications like robot navigation and systems to assist the visually impaired. Caption text or artificial text, on the other hand, is superimposed on videos and, in most cases, is more related to the content. Typical examples of the artificial text include news tickers, anchors' names, scorecards and movie credits etc. The correlation of caption text with the actual content makes it more appropriate for indexing and retrieval applications. An example video frame containing occurrences of the scene as well as caption text is illustrated in Fig. 1.

From the viewpoint of text-based video indexing and retrieval, textual content in video frames needs to be detected (localised) and recognised. Keywords in the transcription of a video frame can then be extracted and employed for indexing and subsequent retrieval. The present study focuses on the latter of these, i.e. recognition of video text formally known as video optical character recognition (V-OCR). In contrast to printed text, recognition of text appearing in video frames is marked by many challenges. Typical problems include low resolution of text, complex and non-homogeneous backgrounds and, different font styles, sizes and colours etc. Another set of challenges is also introduced by the complexity of the script to be recognised. Thanks to more than five decades of extensive research [1–4], mature recognition systems have been developed for recognition of text in non-cursive scripts (languages based on Roman script for instance). Recognition of cursive scripts (like Arabic, Persian, Urdu etc.) is much more challenging and the research attention of the pattern recognition community in this problem is relatively recent (especially for caption text).

To highlight the recognition challenges of cursive script, it is important to mention the complex word formation in such scripts. Typically, a word in a cursive language like Arabic or Urdu is a combination of one or more ligatures where a ligature represents one or more characters joined together through joiner rules. Joiner

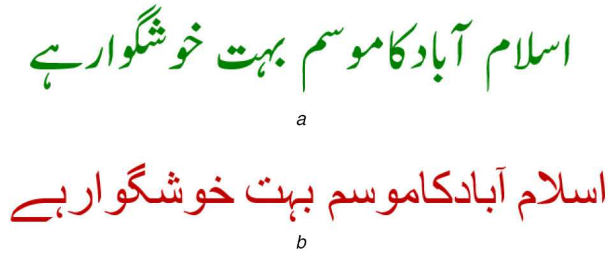

اسلام آباد کا موسم بہت خوشگوار ہے
a
اسلام آباد کا موسم بہت خوشگوار ہے
b

Fig. 2 *Sample text line in*
(a) Nastqliq, ***(b)*** Naskh script

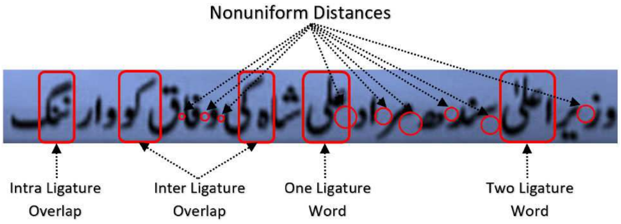


Fig. 3 *Example of cursive text line highlighting recognition challenges*

rules determine which characters are joined and which appear in isolated form. The shape of characters within a ligature is a function of its position (initial, middle, end, isolated etc.). Ligatures can therefore be considered as partial words. Ligatures are further categorised into primary and secondary components, the primary component being the main body of the ligature while the secondary components represent dots and diacritics. It is also worth mentioning that many ligatures share the same primary component and differ only in the number and/or position of dots (leading to high inter-class similarity). Furthermore, the non-uniform intra- and inter-word spacing in such scripts makes segmentation of lines into words highly complex hence ligatures or characters are mostly employed as units of recognition.

Cursive text is printed (or rendered) in one of the standard scripts, Naskh and Nastaliq being two popular scripts (for Arabic and Urdu, respectively). Naskh script follows a horizontal baseline, i.e. characters are joined along a horizontal line. In Nastaliq, on the other hand, characters are joined diagonally making it highly cursive. This diagonal style also results in overlapping of neighbouring characters both horizontally as well as vertically making segmentation of characters much more challenging as compared to Naskh. Fig. 2 illustrates an Urdu text line printed in both Naskh and Nastaliq scripts. Among these, Nastaliq being the more common script for Urdu text makes the subject of our current study. More specifically, we target recognition of cursive Nastaliq text appearing in video frames using Urdu as a case study though the findings can be generalised to other cursive scripts as well. An example cursive (Urdu) text line illustrating various recognition challenges is presented in Fig. 3.

Recognition of cursive text is typically carried out using holistic or analytical methods. Holistic techniques employ ligatures (or words) as recognition units while analytical methods rely on recognition of individual characters either through explicit or implicit segmentation. Holistic techniques avoid the complex character segmentation step but have to deal with a large number of unique ligature classes. On the other hand, in case of analytical techniques, the number of unique classes to recognise is relatively smaller and equals the number of characters in the alphabet (and their various context-dependent shapes). Segmentation of cursive text into characters, however, is a highly challenging problem in itself. The recent advancements in deep neural networks (DNNs) have led to the development of implicit segmentation-based recognition systems where the learning algorithm is provided with images of text lines and the corresponding transcription. The algorithm not only learns character shapes but also the segmentation points hence avoiding the complex explicit segmentation of ligatures into characters.

This paper presents a detailed study on the recognition of cursive caption text. Text line images extracted from video frames are pre-processed and fed to a convolutional neural network (CNN) that extracts a sequence of features using sliding windows. The feature sequences are then used to train a long short-term memory (LSTM) network with a connectionist temporal classification (CTC) layer for sequence alignment. A series of experiments on text lines extracted from News channel videos reports high recognition rates. The key contributions of this study are listed as follows:

- Development of a large data set of video frames along with ground truth information to support the evaluation of text recognition systems.
- Investigation of various pre-processing (binarisation) techniques to segment text from background enhancing the recognition performance.
- A comprehensive recognition framework for cursive caption text using CNN with different variants of recurrent neural networks (RNNs).
- Validation of proposed techniques through an extensive series of experiments.

This paper is organised as follows. In the next section, we present an overview of the current state-of-the-art on recognition of caption text. Section 3 introduces the developed data set along with the labelling tool and ground truth information. The proposed recognition technique is detailed in Section 4 while Section 5 presents the experimental protocol, results and discussion. Finally, we conclude the paper in Section 6 and discuss open challenges and potential research directions on the problem.

2 Related work

OCR is one the most classical pattern recognition problems that have been investigated in images, documents, natural scenes and videos for more than five decades [3]. From the recognition of isolated characters and digits to complex end-to-end systems, the domain has matured significantly over the years. Formally, the task of an OCR system is to take a set of pixel data which contains textual information as input and convert it into a corresponding string as output. Thanks to the extensive research endeavours, mature recognition systems like Google Tessearact [5] and Abbyy FineReader etc. have been developed reporting near to 100% recognition rates on the text in multiple scripts. However, as discussed earlier, recognition of text in cursive scripts still remains challenging especially when it comes to caption text [6].

For text recognition in document images, a number of techniques have been presented both at character (analytical) and word (holistic) levels. Among methods employing characters as units of recognition graph-based models [7–10], Bayesian classifiers [11–13] and hidden Markov models (HMMs) [14–17] etc. have been typically investigated. Among holistic or word level recognition techniques, a wide variety of features as well as classifiers have been studied [18–21] reporting high recognition rates. Deep learning has also been employed for feature extraction and classification at character and word levels [22–24].

Contrary to document images, recognition of text from scene images is much more challenging due to camera perspective, varying lighting conditions and unconstrained backgrounds. Among well-known studies on this problem, histogram of oriented gradients (HOG) [25–27], Strokelets [28, 29], and SIFT descriptors [30, 31] have been proposed as popular features. For classification ANNs [32–34] and SVMs [35, 36] have been widely employed. A special case of scene text recognition is the recognition of text on road signs and has also been investigated in a number of studies [37–41]. Word spotting-based approaches have also been employed on scene text images [42–44] in the literature.

In a number of relatively recent studies, the combination of CNNs with RNN (and variants) have ben effectively applied to text recognition [45–49]. Further advancements in deep learning techniques led to the development of more sophisticated architectures. Notable of these include binary convolutional encoder–decoder network (B-CEDNet) using bidirectional-RNN [50], character-aware neural network (Char-Net) using LSTM [51], character attention fully convolutional network (CA-FCN) [52] and double supervised network (DSAN) [53]. These DNN-based models claim to be robust and efficient for scene text recognition in challenging scenarios.

While the challenges related to camera perspective and non-homogeneous backgrounds are not encountered in case of artificial video text, a major recognition challenge is the possible low resolution of the text. Multiple frame integration technique [54–56] has been widely employed to improve the quality of video image text by generating high resolution images prior to recognition. In [57], a holistic and component-based approach is used for text recognition in low-resolution videos reporting high recognition rates. A new classifier called FCNN (fuzzy-clustering neural network) [58] is proposed for recognition of Chinese text appearing videos. A step-wise language model is identified with ANN to develop new V-OCR in [59]. In another work [60], structural features of characters are employed for recognition of text in videos. Recognition of video text for indexing and retrieval applications has also been explored in a number of studies [61–63].

Among recent studies on this problem, Zayene *et al*. [64] presented a segmentation-free technique based on multidimensional LSTM (MDLSTM) for recognition of Arabic video text. The proposed technique reports high recognition rates on two benchmark data sets, ACTiV [65] and ALIF [66, 67]. Using the same data sets, another study is reported by Jain *et al*. [68] with CNN–LSTM hybrid DNN. Another notable method for video text recognition is presented in [69] where an SVM is employed to select an appropriate colour channel and the selected channel is exploited to recognise text using HMMs. Lu *et al*. [70] employ transfer learning with pre-trained CNNs for video text recognition. State-of-the-art models including InceptionV3, VGG16 and Resnet50 are considered in this study. A similar study for recognition of East Asian languages is presented in [71] where Chinese characters are recognised using CNNs.

Regarding recognition of cursive scripts and more specifically Urdu text, significant research efforts have been made in the last few years to develop Urdu text recognition systems. These systems primarily target scanned document images of printed text (in Nastaliq script). Due to challenges already discussed, implicit segmentation-based techniques have remained a popular choice of researchers [72–75]. Likewise, in the case of holistic approaches, ligatures have been typically employed as recognition units [76].

The initial research endeavours on recognition of Urdu text mainly targeted isolated characters [77, 78] or already segmented ligatures [79]. Among significant holistic approaches, HMMs have been widely employed for recognition of ligatures [80–83]. These techniques use the sliding windows to extract features from ligature images which are projected in the quantised feature space hence representing each ligature image as a sequence. In some cases, the main body and dots are separately recognised [76] to reduce the total number of unique classes which can be very high in case of Urdu text (Urdu has more than 26,000 unique ligatures [84]). A number of holistic techniques are based on word spotting [85, 86] rather than recognition, to retrieve documents containing words similar to those provided as a query.

Among implicit segmentation-based techniques, Hassan *et al*. [87] employ bidirectional LSTM networks with CTC output layer to recognise text from Urdu text lines. In a similar study, Naz *et al*. [72, 88] use MDLSTM with CTC layer and report more than 98% character recognition rate. Most of the studies reported on recognition of Urdu text are evaluated on one of the two publicly available databases, Urdu Printed Text Image (UPTI) [89] and Center for Language Engineering (CLE) [90] data set. UPTI is a synthetically generated data set containing more than 10,000 text lines with ground truth transcription while CLE database contains two parts; the first part contains clusters of high-frequency ligatures while the second part contains scanned images of Urdu books printed in Nastaliq font. A few efforts have also been made towards recognition of Urdu handwritten text [91] and a labelled data set of Urdu handwriting [92] is developed.

The literature is relatively limited when it comes to recognition of Urdu caption text. A holistic approach is presented by Hayat *et al*. [93] where a number of pre-trained CNNs are employed to recognise a small set of 290 unique ligatures. While high recognition rates are reported, the number of ligatures considered is very small. In another recent study [94], bi-directional LSTMs are employed for recognition of Urdu News tickers. The technique is evaluated on a custom developed data set and the performance is compared with a commercial recognition engine.

The discussion on recognition systems presented above summarises the major developments on this problem. The advent of deep learning solutions represents a key milestone which allowed the development of robust recognition systems significantly reducing the error rates. The solution we adopt for recognition of cursive video text is also inspired by the success of different DNNs on similar problems. Prior to presenting the details of the proposed solution, in the next section, we introduce the data set that has been developed as a part of our study.

3 Data set

For algorithmic development and evaluation of the proposed system, we developed a benchmark data set of video frames that are labelled from two perspectives, detection of text regions and recognition of text. Detection performance refers to how effective the system is in localising the textual content while recognition performance evaluates the effectiveness of converting images into text. From the viewpoint of localisation, the bounding box of each textual region in the frame is identified and stored. Similarly, for recognition performance evaluation, transcription of each textual region needs to be stored in ground truth information. In the present study, we focus on the recognition task hence the textual regions are extracted from video frames using the ground truth information. The extracted text regions are then employed to train and evaluate the system.

We have collected 46 videos from different News channels, all videos are recorded at a frame rate of 25 fps with a resolution of 900 × 600. Due to redundant information in successive frames, one frame every 2 s is extracted and labelled. Samples of the frames from different News channels are presented in Fig. 4 while important statistics on collected videos, frames, text lines, words and characters are summarised in Table 1.

In order to facilitate the ground truth labelling process, a labelling tool (Fig. 5) was developed that allows storing the location and ground truth transcription of each textual region in a video frame. The ground truth information associated with a frame is stored in the XML format. The ground truth file consists of two parts, frame meta-data and information of the textual content. Meta-data of each frame contains information on video, channel ID and a unique code for identifying the frames. The second part of the XML file has information of each text region in the frame along with a unique identity for each and, type of text (scene text or caption text). Localisation information of a text region is stored with its x and y coordinates and the *width* and *height* of the bounding rectangle. Transcription of text is stored using Unicodes of each character. Fig. 6 illustrates the ground truth information of an example frame stored in XML file.

As mentioned earlier, text lines extracted from video frames using the ground truth information are used to train the learning algorithm. Few examples of text lines are illustrated in Fig. 7. In an attempt to enhance the size of training data (to ensure maximum representation of various character shapes and their combinations), we also generated a set of 50,000 synthetic text lines. The lines are generated using the text taken from different books and News portals. To ensure a close resemblance with the actual data, various backgrounds are extracted from actual News channel videos and the synthetically generated text is superimposed on these backgrounds (Fig. 8). Samples of such synthetic text lines are presented in Fig. 9 where it can be seen that the generated text line images look very similar to the actual text lines extracted from video frames.

4 Methods

This section presents the details of the proposed recognition technique. The text line images are first pre-processed to segment

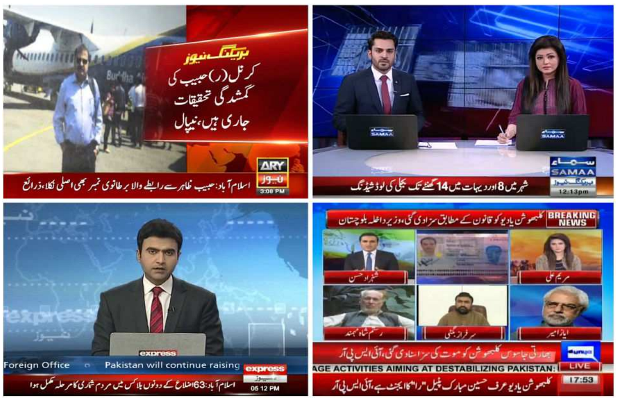
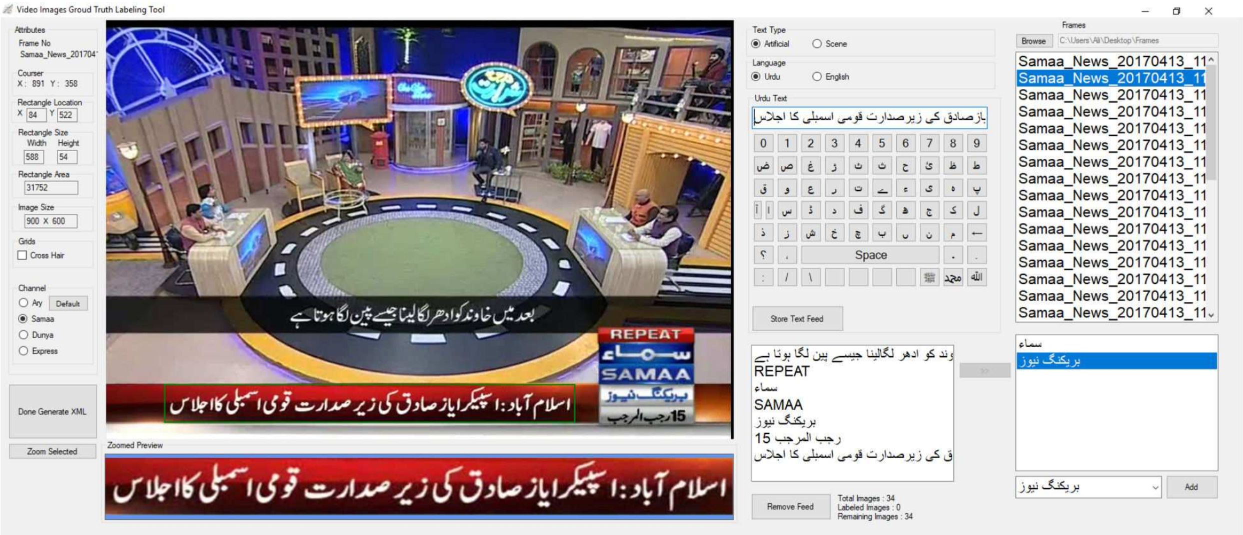


Fig. 4 *Sample video frames from different News channels*

Table 1 Statistics of labelled video frames

S#	Channel	Videos	Labelled images	Urdu lines	English lines
1	Ary news	7	3206	10,250	3605
2	Samaa news	13	2503	10,961	4411
3	Dunya news	16	3059	10,723	8,861
4	Express news	10	2424	8536	6755
	total	46	11,192	40,470	23,632



Video Images Groud Truth Labeling Tool

Attributes
Frame No Samaa_News_201704
Courser X: 891 Y: 358
Rectangle Location X 84 Y 522
Rectangle Size Width 588 Height 54
Rectangle Area 31752
Image Size 900 X 600
Grids ☐ Cross Hair
Channel ○ Ary ● Samaa ○ Dunya ○ Express

Text Type ● Artificial ○ Scene
Language ● Urdu ○ English
Urdu Text: ﺎﺯﺻﺎﺩﻕ ﮐﯽ ﺯﯾﺮﺻﺪﺍﺭﺕ ﻗﻮﻣﯽ ﺍﺳﻤﺒﻠﯽ ﮐﺎ ﺍﺟﻼﺱ

Frames: C:\Users\Ali\Desktop\Frames

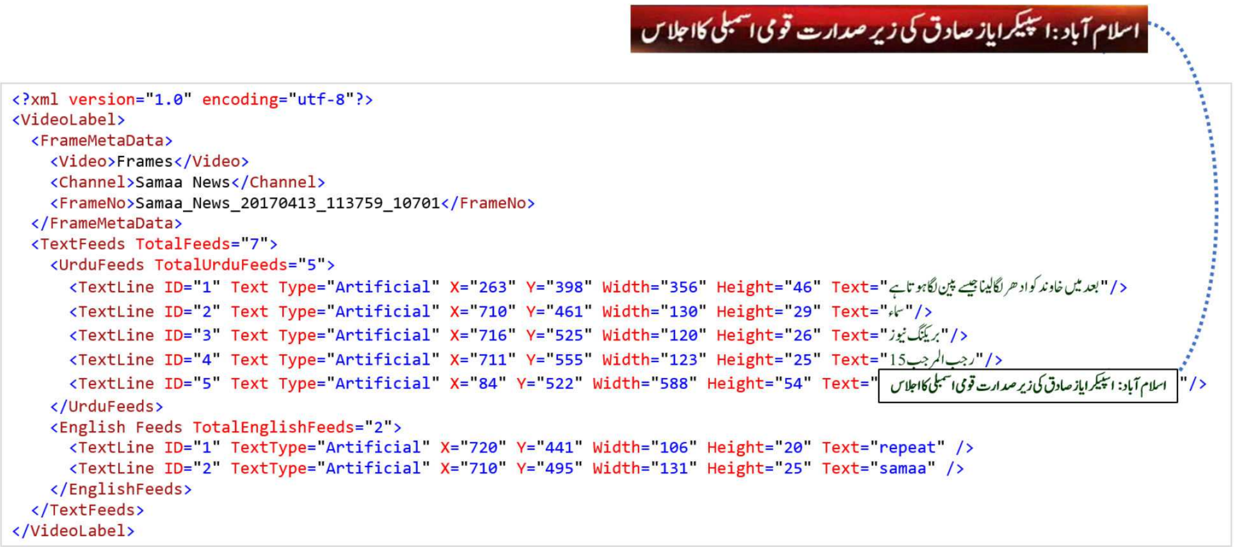
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11
Samaa_News_20170413_11

وند کو ادھر لگالینا جیسے پین لگا ہوتا ہے
REPEAT
سماء
SAMAA
بریکنگ نیوز
15 رجب المرجب
ق کی زیرصدارت قومی اسمبلی کا اجلاس

سماء
بریکنگ نیوز

Total Images : 34
Labeled Images : 0
Remaining Images : 34

Fig. 5 *Ground truth labelling tool*



```
<?xml version="1.0" encoding="utf-8"?>
<VideoLabel>
  <FrameMetaData>
    <Video>Frames</Video>
    <Channel>Samaa News</Channel>
    <FrameNo>Samaa_News_20170413_113759_10701</FrameNo>
  </FrameMetaData>
  <TextFeeds TotalFeeds="7">
    <UrduFeeds TotalUrduFeeds="5">
      <TextLine ID="1" Text Type="Artificial" X="263" Y="398" Width="356" Height="46" Text="بعد میں خاوند کو ادھر لگالینا جیسے پین لگا ہوتا ہے"/>
      <TextLine ID="2" Text Type="Artificial" X="710" Y="461" Width="130" Height="29" Text="سماء"/>
      <TextLine ID="3" Text Type="Artificial" X="716" Y="525" Width="120" Height="26" Text="بریکنگ نیوز"/>
      <TextLine ID="4" Text Type="Artificial" X="711" Y="555" Width="123" Height="25" Text="15رجب المرجب"/>
      <TextLine ID="5" Text Type="Artificial" X="84" Y="522" Width="588" Height="54" Text="اسلام آباد: اسپیکر ایاز صادق کی زیر صدارت قومی اسمبلی کا اجلاس"/>
    </UrduFeeds>
    <English Feeds TotalEnglishFeeds="2">
      <TextLine ID="1" TextType="Artificial" X="720" Y="441" Width="106" Height="20" Text="repeat" />
      <TextLine ID="2" TextType="Artificial" X="710" Y="495" Width="131" Height="25" Text="samaa" />
    </EnglishFeeds>
  </TextFeeds>
</VideoLabel>
```

Fig. 6 *Labelled frame information in an XML file containing bounded-boxes and transcriptions of text*

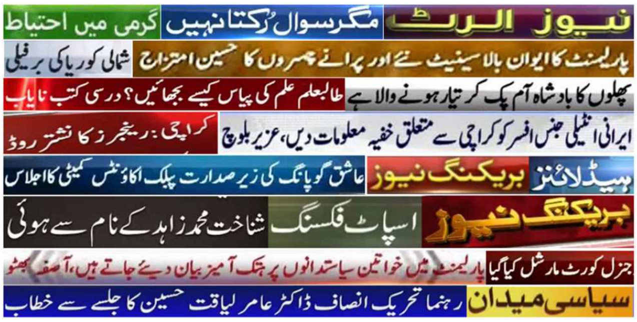


Fig. 7 *Sample text lines extracted from video frames*

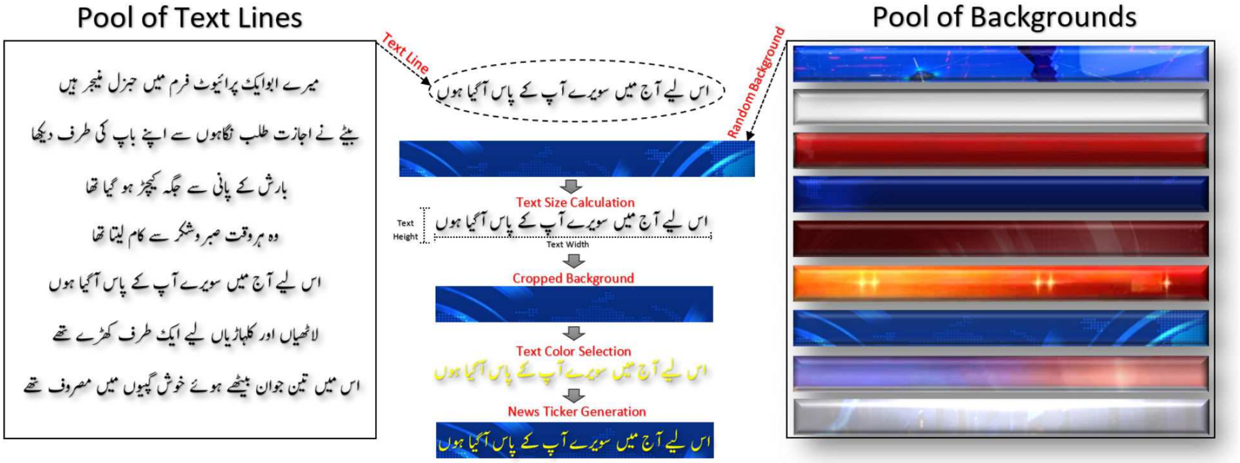

Pool of Text Lines
میرے ابوایک پرائیوٹ فرم میں جنرل منیجر ہیں
بیٹے نے اجازت طلب نگاہوں سے اپنے باپ کی طرف دیکھا
بارش کے پانی سے جگہ کیچڑ ہو گیا تھا
وہ ہر وقت صبر و شکر سے کام لیتا تھا
اس لیے آج میں سویرے آپ کے پاس آگیا ہوں
لاٹھیاں اور کلہاڑیاں لیے ایک طرف کھڑے تھے
اس میں تین جوان بیٹھے ہوئے خوش گپیوں میں مصروف تھے
Text Line
Random Background
Text Size Calculation
Text Height
Text Width
Cropped Background
Text Color Selection
News Ticker Generation
Pool of Backgrounds

Fig. 8 *Generation of synthetic text lines*

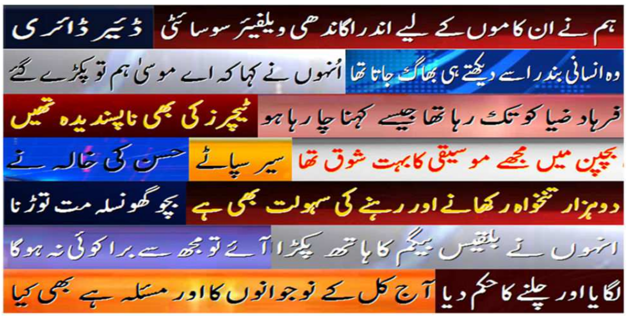


Fig. 9 *Examples of synthetically generated text lines*

text from the background. The binarised images of text lines are then fed to a CNN for feature extraction. The generated feature map is then fed to an RNN using sliding windows. Finally, being a sequence-to-sequence mapping problem, we need a CTC layer for sequence alignment. An overview of these steps is presented in Fig. 10 while each of these is detailed in the following.

4.1 Pre-processing

While the recognition engine can be fed with coloured or greyscale images, removing the background information and binarising the image allows the learning algorithm better learn character shapes and boundaries. For images with simple and homogenous backgrounds, global thresholding suffices. Video frames, however, often contain text on multiple, non-homogeneous backgrounds. Furthermore, there are two scenarios in which text may appear; dark text on a bright background or bright text on dark background. Once the image is binarised, we need all text lines to follow one of the two conventions. In our study, we assume dark text on a bright background and if this is not the case, we invert the polarity of the greyscale image prior to binarisation.

As a first step, we need to detect the polarity of the text. The canny edge detector is applied to the greyscale text line image and blobs are identified. These blobs correspond to text regions in the image. Region filling is applied to these blobs and the generated binary image is used as a mask on the greyscale image to extract potential text regions (characters or ligatures). We then compute the median grey value (Med_{text}) of the extracted blobs as well as the median grey value of the background (all pixels which do not belong to any blob), Med_{back}. If $\text{Med}_{\text{text}} < \text{Med}_{\text{back}}$ we have dark text on bright background and the polarity agrees with our assumed convention. On the other hand, if $\text{Med}_{\text{text}} > \text{Med}_{\text{back}}$, this corresponds to bright text on dark background. In such cases, the polarity of the image is reversed prior to any further processing. The process is summarised in Fig. 11 while more details can be found in our previous work [95].

Once all text lines contain text in the same polarity, we binarise the images to contain only textual information. For binarisation, we investigated a number of thresholding techniques. These include Otsu's global thresholding method [96] as well as a number of local thresholding algorithms. The local thresholding algorithms are adaptive techniques where the threshold value of each pixel is computed as a function of the neighbouring pixels (rather than empirically fixing the threshold). Most of these algorithms are inspired by the classical Niblack thresholding [97] where the threshold is computed as a function of the mean and standard deviation of the grey values in the neighbourhood of a reference pixel. Other algorithms investigated in our study include Sauvola

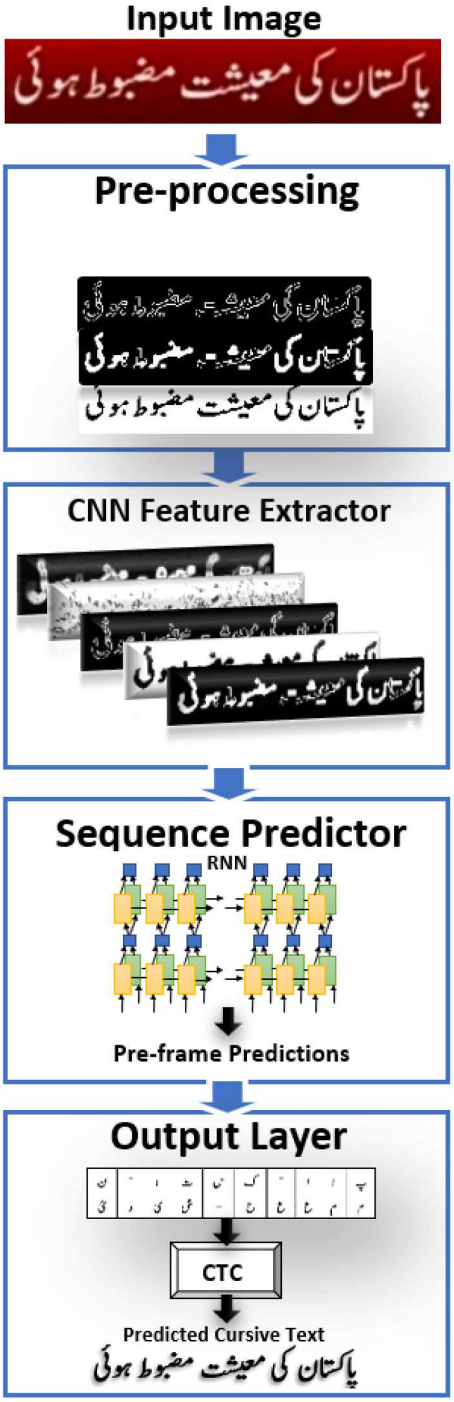


Fig. 10 *Overview of the key processing steps*

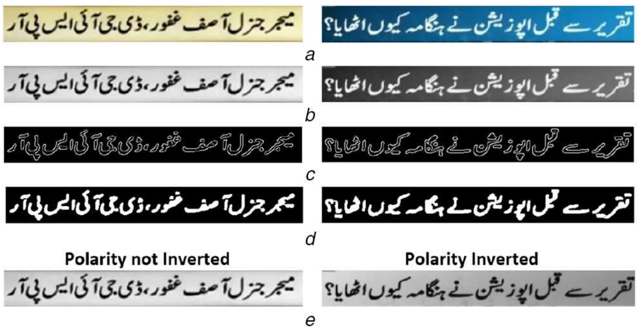

میجر جنرل آصف غفور، ڈی جی آئی ایس پی آر
تقریر سے قبل اپوزیشن نے ہنگامہ کیوں اٹھایا؟
a
b
c
d
Polarity not Inverted
Polarity Inverted
e

Fig. 11 *Identification of polarity of text*
(a) Original image, ***(b)*** Grey scale image, ***(c)*** Text blobs, ***(d)*** Filled text blobs serving as a mask to extract corresponding blobs from the grey image, ***(e)*** Final image (Image on the right is inverted while the one on left remains unchanged)

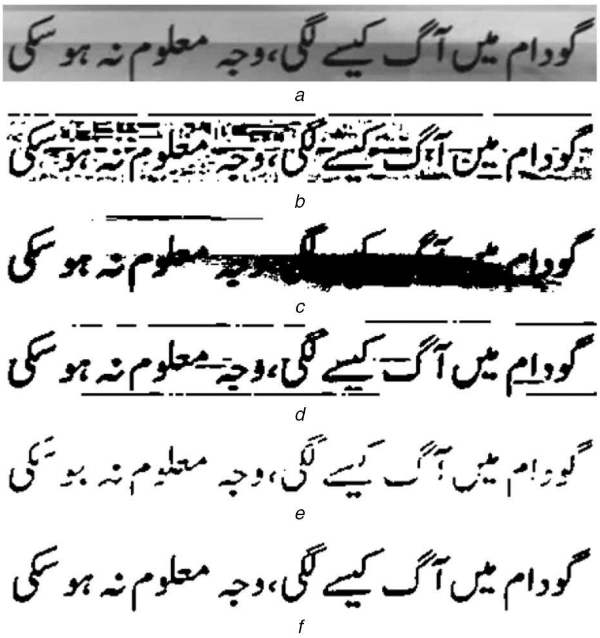


Fig. 12 *Binarisation results on a sample text line*
(a) Greyscale image, ***(b)*** Niblack, ***(c)*** Ostu's global thresholding, ***(d)*** Feng's algorithm, ***(e)*** Sauvola, ***(f)*** Wolf's algorithm

[98], Feng's [99] and Wolf's thresholding algorithm [100]. Prior to binarising the images, we also apply a smoothing (median) filter on each text line to remove/suppress any noisy patterns in the image. Binarisation results of applying various thresholding techniques to a sample text line image are illustrated in Fig. 12. From the subjective analysis of these results, Wolf's algorithm that was specifically proposed for low-resolution video text seems to outperform other techniques. Nevertheless, it is hard to generalise from visual inspection of a few sample images and the recognition rates on images generated by each of these techniques could be a better indicator of the effectiveness of the method. Following binarisation, we normalise the height of each text line to a fixed size (90 pixels in our case) while the width of the line is a function of the textual content it contains.

4.2 Feature extraction

Once the text lines are pre-processed, we proceed to the next step of feature extraction. As mentioned earlier, the text line image contains a sequence of characters which needs to be mapped to the corresponding sequence of characters in the ground truth transcription. The problem is hence formulated in a sequence-to-sequence mapping framework. The input sequences to the model can be raw pixel values or features (hand-crafted or machine-learned) extracted using a sliding window protocol. A number of recent studies [101–103], validated the superiority of machine-learned features over hand-engineered features (and raw pixel values). We, therefore, employ a CNN as a feature extractor. The architecture of CNN is a function of many hyper-parameters. In our study, we have employed seven convolutional layers (with max-pooling). The output of CNN is a 128×512 dimensional feature vector. The architecture of the employed CNN can be seen in Fig. 13 while Fig. 14 demonstrates the correspondence between features and the respective regions in the image.

4.3 Sequence predictor

Feature sequences computed using CNN are fed to a sequence predictor for recognition. More specifically, we employed a bidirectional (RNN) and its different variants for sequence recognition. A simple RNN cell works by multiplying current input x_t and previous output h_{t-1} and employs the tanh activation function. Such vanilla RNNs are known to suffer from the vanishing (exploding) gradient problem during training hence different variants of simple RNNs were introduced. Common of

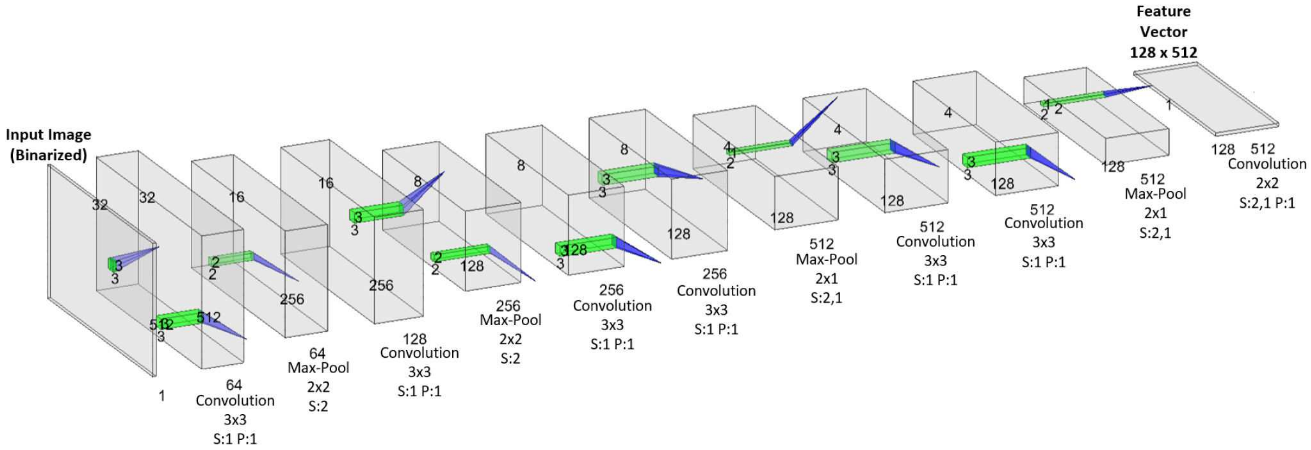


Fig. 13 *Architecture of the CNN employed for feature extraction*

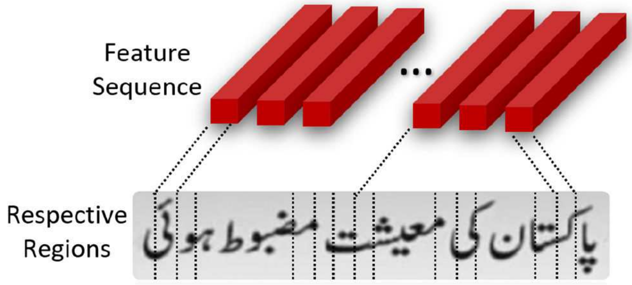


Fig. 14 *Feature extraction form text line image; each vector is associated with the respective region of the input image*

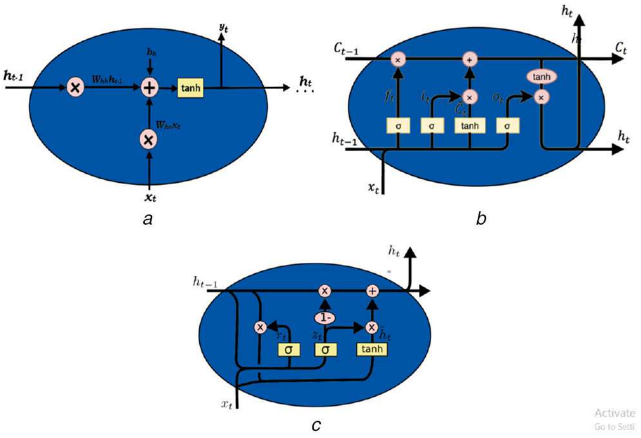


Fig. 15 *Architectures of*
(a) Simple RNN, ***(b)*** GRU, ***(c)*** LSTM

these include the LSTM networks and gated recurrent units (GRUs). These sophisticated variants of RNNs are able to model long term dependencies in the input using gates. An LSTM has three (input, forget and output) while a GRU has two (reset and update) gates. Fig. 15 illustrates the architectures of simple RNN, GRU and LSTM cells. The proposed RNN architecture contains 512 time-steps as input and 2 stacks of hidden layers. Each stack contains a forward and a backward layer with 256 hidden units. Fig. 16 shows the employed model which produces a sequence of characters as output. These predictions are then passed to a CTC layer (discussed in the next section) for text alignment.

4.4 CTC layer

A CTC layer serves to convert the raw predictions of RNN into the actual transcription of a given text line image aligning the output sequence of RNN with the target labels. The alignment of labels is learned during the training process. The CTC layer keeps a record of all labels in the transcription along with an extra character which separates the consecutive occurrence of characters in transcription. CTC layer predicts the most probable sequence of labels against the sequence predicted by the RNN (Fig. 16).

5 Experiments and results

To evaluate the effectiveness of the proposed recognition technique, we carried out an extensive series of experiments. We first introduce the experimental protocol followed by the recognition results as a function of pre-processing, type of RNN cell and various combinations of training data. Finally, we present a comparative analysis of the reported results with respect to other similar studies.

5.1 Experimental protocol

The experimental study of the system is carried out on text lines extracted from the video frames using ground truth information (Section 3). The total number of text lines extracted from video frames of four different News channels sum up to a total of 40,470. Among these, 27,321 text lines are used in the training set, 4000 text lines (1000 from each channel) in the validation set while 9149

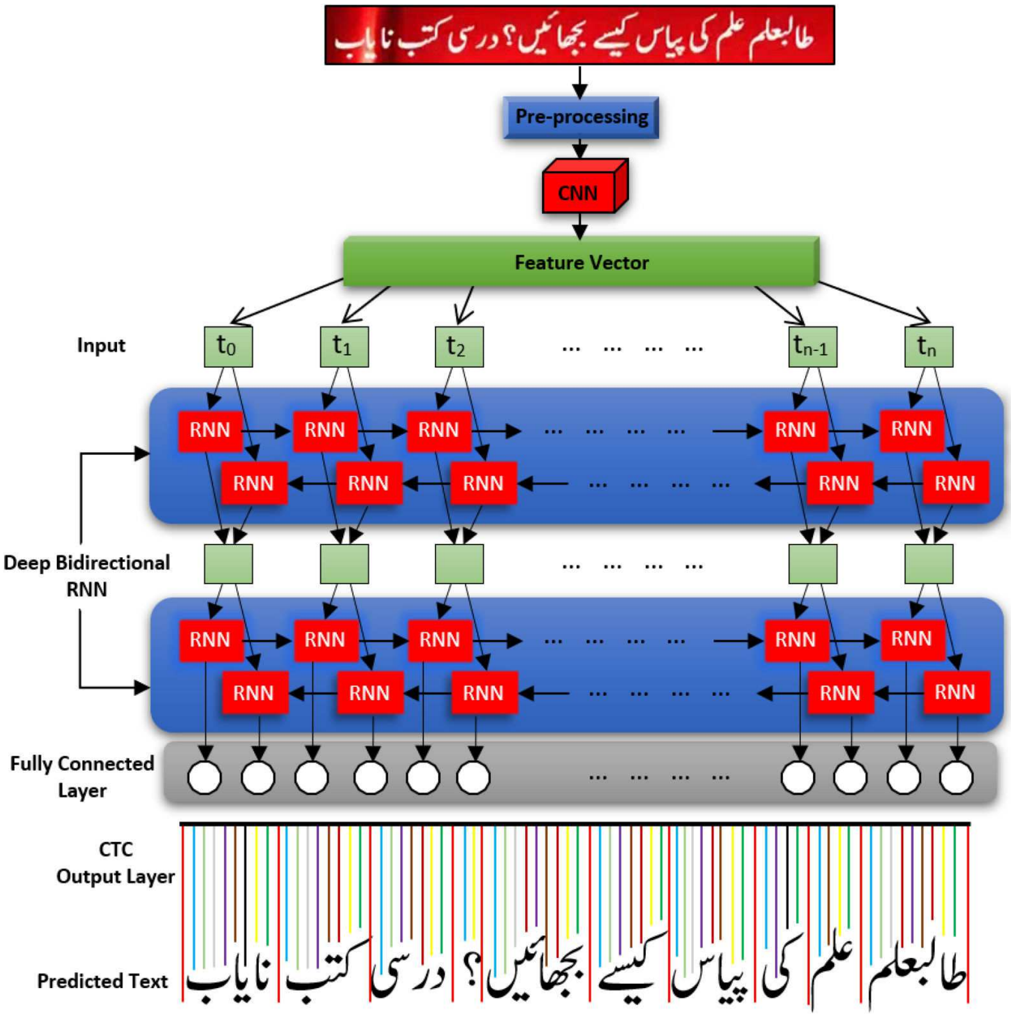


Fig. 16 *Architecture of the bidirectional (left-to-right & right-to-left) RNN model with CTC output layer*

Table 2 Distribution of text lines in training, validation and test sets

S#	Channel	Training	Validation	Test
1	Ary news	7740	1000	1510
2	Dunya news	6702	1000	3021
3	Express news	5120	1000	2416
4	Samaa news	7759	1000	2202
	total	27,321	4000	9149

Table 3 Distribution of the data set including synthetically generated text lines

	Train		Validation		Test	
type of data	lines	chars	lines	chars	lines	chars
video text	27,321	556,773	4000	81,516	9149	277,819
synthetic text	50,000	1,259,339	—	—	—	—
total	77,321	1,816,112	4000	81,516	9149	277,819

text lines are used in the test set. The distribution of text lines into training, validation and test sets is summarised in Table 2. In some of the experiments, we also employed 50,000 synthetic text lines. These synthetic text lines, however, are only employed in the training set (Table 3) while the validation and test sets for all experiments are kept the same. The validation set is employed to chose the different hyper-parameters of the recogniser.

5.2 Recognition results

In the first series of experiments, we studied the recognition performance as a function of the pre-processing (binarisation) technique and the type of RNN cell (simple RNN, GRU, and LSTM). These experiments are carried out on actual text lines only and synthetic data is not included at this stage, i.e. 27,321 text lines in the training set, 4000 in validation and 9149 in the test set (Table 2). Furthermore, we also compared the performance of feeding the RNN with raw pixel values and CNN-based features. For all experiments, we quantify the system performance by computing the character recognition rates. The recognition engine outputs the predicted transcription of the query text line. Recognition rates are calculated by computing the Levenshtein distance between the predicted and the ground truth transcription. The recognition rates corresponding to the first series of experiments are summarised in Table 4.

A number of interesting observations can be made from the reported recognition rates. First of all, it can be seen that features computed using CNN report higher recognition rates in all

Table 4 Summary of character recognition rates in different experimental settings

	Character recognition rate, %					
		Raw pixels			CNN features	
	RNN	GRU	LSTM	RNN	GRU	LSTM
Greyscale	72.57	75.19	77.68	80.09	82.35	84.1
Niblack [97]	70.66	73.39	74.81	76.87	79.91	82.13
Otsu [96]	69.36	71.45	75.23	76.54	79.11	80.39
Feng [99]	76.67	79.85	81.31	82.94	84.56	87.91
Sauvola [98]	74.27	78.36	80.19	80.65	83.41	86.25
Wolf [100]	81.78	83.88	86.06	90.18	92.76	95.98

Table 5 Recognition rates as a function of training data

Item	Recognition rate, %	
Training data	Line	Chars
videos	77.53	95.98
synthetic	69.68	89.32
videos + synthetic	81.34	97.63

experiments as compared to raw pixels. Secondly, RNNs implemented with GRU perform better than simple RNN cells while LSTM-based model outperforms the other two in all cases. Comparing the various binarisation techniques, the greyscale text lines report higher recognition rates when compared to those obtained on text lines binarised using Niblack and Otsu's thresholding algorithms. This observation is consistent with our initial assessment of binarisation algorithms where, in general, Niblack's binarisation introduces a lot of noise in the binarised images while global thresholding fails once the text images have non-homogeneous backgrounds. The performance of Feng's and Sauvola's binarisation methods is more or less similar. Text lines binarised using Wolf's algorithm report the highest recognition rates. This observation is also consistent with the subjective analysis of binarisation techniques where Wolf's algorithm produced relatively cleaner versions of binarised images. Overall, the highest reported recognition rate reads 95.98% when using the CNN–LSTM combination and binarising the text lines using Wolf's algorithm. Based on these observations, the subsequent experiments are carried out with Wolf's binarisation technique as the pre-processing step and the combination of CNN and LSTM as the recognition model.

In the second series of experiments, we study the impact of training data on the recognition performance (using CNN–LSTM with Wolf's binarisation). Furthermore, to provide deeper insights, in addition to character recognition rates, we also computed the line recognition rate. A text line is considered to be correctly recognised once all characters within the line are recognised correctly. The models are trained using three different scenarios, using text lines from video frames, using synthetic text lines only and by combining the video text lines with synthetically generated text lines. The results of these experiments are presented in Table 5. It is interesting to note that when the system is trained using only synthetic data, it still reports acceptable recognition rates reading 69.68 and 89.32% at the line and character levels, respectively. Combining the video text lines with synthetic text lines improves the character recognition rate from 95.98 to 97.63% demonstrating the effectiveness of the generated text lines. The enhanced recognition rates when using synthetic data can be attributed to the fact that some of the character combinations which could not be captured in the original text lines are represented in the synthetic text lines leading to improved recognition rates. Training with synthetic data, naturally, took slightly longer to converge (Fig. 17) as the learning algorithm has more number of character combinations to learn.

In the last series of experiments, we studied the impact of the size of training data on the recognition performance. Keeping the test size (and all other system parameters) fixed, we varied the number of training text lines from 3000 to 27,321. The corresponding recognition rates are illustrated in Fig. 18 where it can be seen that the recognition rates being to stabilise from 15,000 lines of text onwards which is a manageable size for such applications.

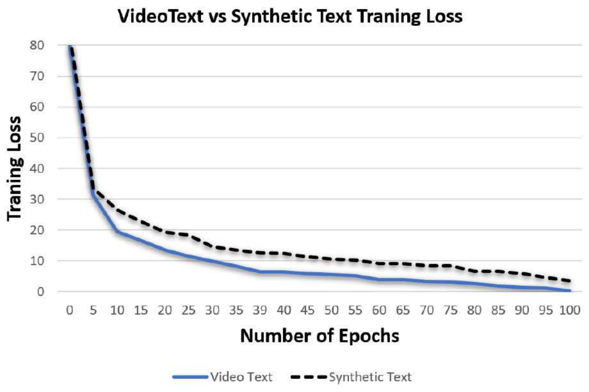


Fig. 17 *Training loss for video and synthetically generated text lines*

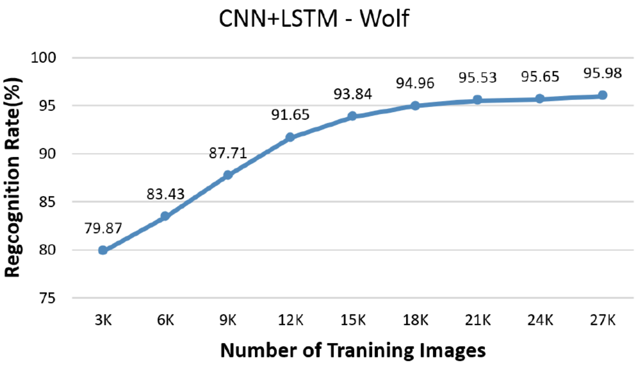


Fig. 18 *Recognition rates as a function of the size of training data*

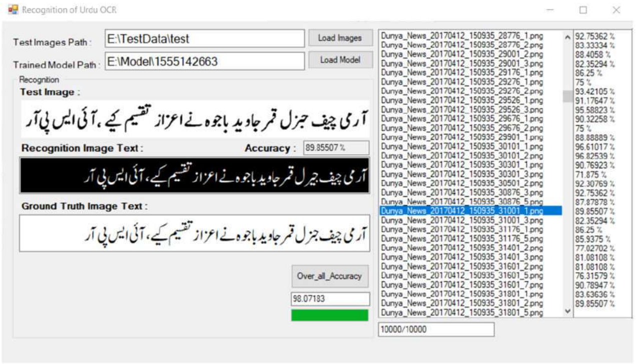
From the viewpoint of recognition time, the time to recognise a text line is naturally a function of the length of text. We report the average recognition time per line, the average being computed on all text lines in our test set. Recognition takes on average 0.18 s per text line on Tesla K40 GPU Computing Processor with 12GB RAM.Video frames, on the average, contain four to five text lines hence the recognition engine can process one frame per second allowing it to be employed for indexing and retrieval applications.

5.3 Performance comparison

To provide an idea of the effectiveness of the proposed recognition technique, we present a comparative analysis of various recent studies. Naturally, a meaningful quantitative comparison is only possible if all techniques are evaluated on the same data set using the same experimental protocol. However, unfortunately, due to the

Table 6 Performance comparison with other recognition techniques

Image type	Study	Language	Technique	Database	Data size	Results, %
document	Ahmed *et al.* [104]	Urdu	ANN	Private	56 character clusters	93.40
	Akram *et al.* [105]	Urdu	DCT & HMMs	CLE	224 images	86.15
	Hussain *et al.* [106]	Urdu	DCT & HMMs	CLE	5,249 ligatures	87.44
	Ahmed *et al.* [107]	Urdu	Raw-Pixels & BLSTM	UPTI	15,251 lines	96.00
	Naz *et al.* [108]	Urdu	Stat Features & MDRNN	UPTI	10,000 lines	94.97
	Naz *et al.* [72]	Urdu	Stat Features & MDRNN	UPTI	10,000 lines	96.40
	Naz *et al.* [74]	Urdu	CNN & MDRNN	UPTI	10,000 lines	98.12
	Hassan *et al.* [87]	Urdu	Raw-Pixels with BLSTM	UPTI	10,000 lines	94.85
Videos	Zayene *et al.* [64]	Arabic	MDLSTM	AcTiV-R	7,843 lines	96.85
	Tayyab *et al.* [94]	Urdu	CRNN	Private	19,824 lines	93.02
	Hayat *et al.* [93]	Urdu	CNN	Private	290 uniqe ligatures	99.50
	proposed	Urdu	CNN+LSTM	Private	40,470 lines	95.98 (97.63)



Test Images Path : E:\TestData\test

Trained Model Path : E:\Model\1555142663

Recognition

Test Image :

آرمی چیف جنرل قمر جاوید باجوہ نے اعزاز تقسیم کیے، آئی ایس پی آر

Recognition Image Text : Accuracy : 89.85507 %

آرمی چیف جیرل قمر جاوید باجوہ نے اعزاز تقسیم کیے، آئی ایس پی آر

Ground Truth Image Text :

آرمی چیف جنرل قمر جاوید باجوہ نے اعزاز تقسیم کیے، آئی ایس پی آر

98.07183

Dunya_News_20170412_150935_28776_1.png	92.75362 %
Dunya_News_20170412_150935_28776_2.png	83.33334 %
Dunya_News_20170412_150935_29001_2.png	88.4058 %
Dunya_News_20170412_150935_29001_3.png	82.35294 %
Dunya_News_20170412_150935_29176_1.png	86.25 %
Dunya_News_20170412_150935_29276_1.png	75 %
Dunya_News_20170412_150935_29276_2.png	93.42105 %
Dunya_News_20170412_150935_29526_1.png	91.17647 %
Dunya_News_20170412_150935_29526_3.png	95.58823 %
Dunya_News_20170412_150935_29676_1.png	90.32258 %
Dunya_News_20170412_150935_29676_2.png	75 %
Dunya_News_20170412_150935_29901_1.png	88.88889 %
Dunya_News_20170412_150935_30101_1.png	96.61017 %
Dunya_News_20170412_150935_30101_2.png	96.82539 %
Dunya_News_20170412_150935_30301_1.png	90.76923 %
Dunya_News_20170412_150935_30301_3.png	71.875 %
Dunya_News_20170412_150935_30501_2.png	92.30769 %
Dunya_News_20170412_150935_30876_3.png	92.75362 %
Dunya_News_20170412_150935_30876_5.png	87.87878 %
Dunya_News_20170412_150935_31001_1.png	89.85507 %
Dunya_News_20170412_150935_31001_3.png	82.35294 %
Dunya_News_20170412_150935_31176_1.png	86.25 %
Dunya_News_20170412_150935_31176_5.png	85.9375 %
Dunya_News_20170412_150935_31401_2.png	77.02702 %
Dunya_News_20170412_150935_31401_3.png	81.08108 %
Dunya_News_20170412_150935_31601_2.png	81.08108 %
Dunya_News_20170412_150935_31601_5.png	76.31579 %
Dunya_News_20170412_150935_31601_7.png	90.78947 %
Dunya_News_20170412_150935_31801_1.png	83.63636 %
Dunya_News_20170412_150935_31801_2.png	89.85507 %
Dunya_News_20170412_150935_31801_5.png	

10000/10000

Fig. 19 *Screen shot of the recognition application developed in C#.NET and Python*

lack of benchmark data sets for this problem, the reported techniques are mostly evaluated on the custom developed data set. Nevertheless, for completeness, we present these results to give readers an idea of the current state-of-the-art on this problem. Furthermore, in addition to caption text, we also list the recognition rates reported on printed document text in well-known recent studies. These results are summarised in Table 6 with a summary of techniques and the size of data set employed. In the case of printed text, the highest reported recognition rate is 98.12% on the UPTI data set. It is, however, important to mention that UPTI is a synthetically generated data set that does not offer the same kind of challenges as those encountered in scanned images of documents. In the case of caption text, a recognition rate of 96.85% is reported on a relatively smaller set of Arabic text lines. For Urdu caption text, Tayyab *et al.* [94] achieve 93% recognition rate on approximately 20,000 text lines while Hayyat *et al.* [93] report a ligature recognition rate of 99.5%. The data set considered in [93] however, is fairly limited with only 290 unique ligature classes. In our experiments, we report a recognition rate of 95.98% (97.63% with synthetic data in training) which, though not directly comparable with reported studies, is indeed very promising considering the complexity of the problem.

Fig. 19 presents a screenshot of the visual application that was developed for recognition of text lines. Furthermore, to provide insights into recognition errors, some common errors produced by the system are illustrated in Fig. 20 where it can be seen that a major proportion of errors results due to false recognition of secondary ligatures (dots and diacritics) while the main body ligatures is correctly recognised in most cases.

6 Conclusion

We presented an effective technique for recognition of cursive caption text appearing in video frames. The technique relies on a deep learning-based framework and exploits a combination of CNN and RNN for recognition. A comprehensive data set of video frames was developed and labelled as a part of this study. Text line images were extracted from the video frames using ground truth information. The extracted lines were preprocessed and fed to the CNN–RNN model for end-to-end training. A detailed series of experiments was carried out to study the recognition performance with respect to the type of pre-processing applied, type of RNN cells and training data. The system reported a high recognition rate of 97.63% on more than 9000 text line images.

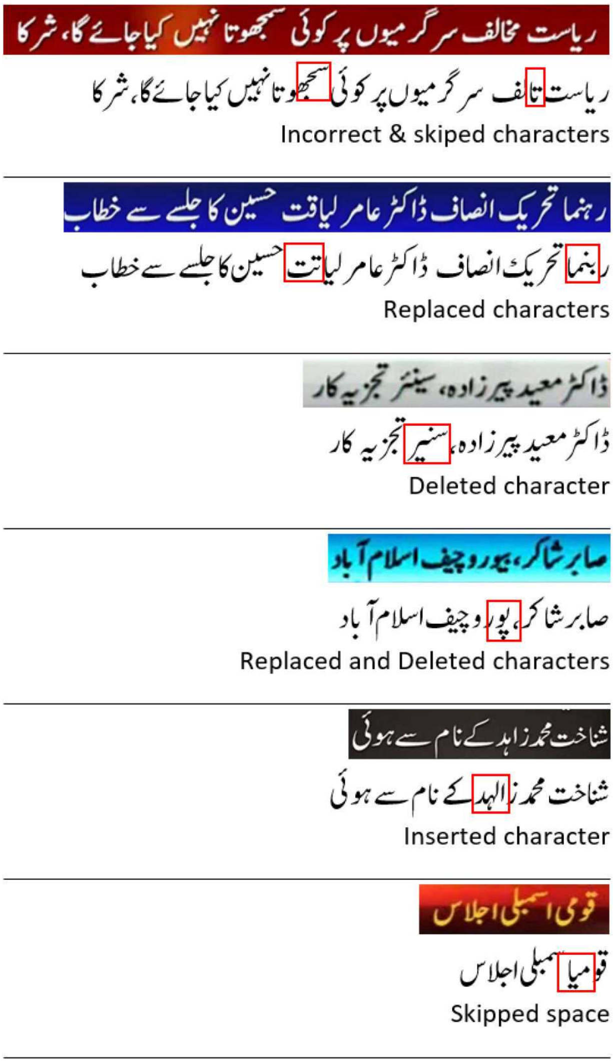

ریاست مخالف سرگرمیوں پر کوئی سمجھوتا نہیں کیا جائے گا، شرکا
Incorrect & skiped characters
رہنما تحریک انصاف ڈاکٹر عامر لیاقت حسین کا جلسے سے خطاب
Replaced characters
ڈاکٹر معید پیرزادہ، سینئر تجزیہ کار
Deleted character
صابر شاکر، بیورو چیف اسلام آباد
Replaced and Deleted characters
شناخت محمد زاہد کے نام سے ہوئی
Inserted character
قومی اسمبلی اجلاس
Skipped space

Fig. 20 *Examples of recognition errors*

In our further study, we intend to combine the recognition engine with our text detection (and localisation) module and evaluate the combined detector and recogniser in an end-to-end fashion. Furthermore, the system will be extended to a complete indexing and retrieval framework where users can provide keywords as query and the system is able to retrieve all videos where the keyword has appeared. In addition to text, we are also investigating video retrieval using spoken keywords as well as visual objects of interest like persons or building etc.

7 Acknowledgments

The authors thank the IGNITE, National Technology Fund, Pakistan for supporting this study under grant no. ICTRDF/TR&D/2014/35.

8 References

[1] Ciardiello, G., Scafuro, G., Degrandi, M., *et al.*: 'An experimental system for office document handling and text recognition'. Proc 9th Int. Conf. on Pattern Recognition, Rome, Italy, 1988, pp. 739–743

[2] Elliman, D.G., Lancaster, I.T.: 'A review of segmentation and contextual analysis techniques for text recognition', *Pattern Recognit.*, 1990, **23**, (3–4), pp. 337–346

[3] Qiaoyang, Y., Doermann, D.: 'Text detection and recognition in imagery: a survey', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (7), pp. 1480–1500

[4] Komanduri, S., Roopa, Y.M., Bala, M.M.: 'Novel approach for image text recognition and translation'. 2019 3rd Int. Conf. on Computing Methodologies and Communication (ICCMC), Tamil Nadu, India, 2019, pp. 596–599

[5] Smith, R.: 'An overview of the tesseract OCR engine'. Ninth Int. Conf. on Document Analysis and Recognition, 2007. ICDAR 2007, Curitiba, Brazil, 2007, vol. 2, pp. 629–633

[6] Naz, S., Hayat, K., Razzak, M.I., *et al.*: 'The optical character recognition of Urdu-like cursive scripts', *Pattern Recognit.*, 2014, **47**, (3), pp. 1229–1248

[7] Palaiahnakote, S., Quy, P.T., Lim, T.C.: 'A Laplacian approach to multi-oriented text detection in video', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (2), pp. 412–419

[8] Garz, A., Seuret, M., Simistira, F., *et al.*: 'Creating ground truth for historical manuscripts with document graphs and scribbling interaction'. 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 2016, pp. 126–131

[9] Toselli, A.H., Vidal, E., Romero, V., *et al.*: 'HMM word graph based keyword spotting in handwritten document images', *Inf. Sci.*, 2016, **370**, pp. 497–518

[10] Fasquel, J.-B., Delanoue, N.: 'A graph based image interpretation method using a priori qualitative inclusion and photometric relationships', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, **41**, (5), pp. 1043–1055

[11] Weinman, J.J., Learned-Miller, E.: 'Improving recognition of novel input with similarity'. 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, New York, NY, USA, 2006, vol. 1, pp. 308–315

[12] Islam, N., Islam, Z., Noor, N.: 'A survey on optical character recognition system', arXiv preprint arXiv:1710.05703

[13] Lei, B., Xu, G., Feng, M., *et al.*: '*Classification, parameter estimation and state estimation: an engineering approach using MATLAB*' (John Wiley & Sons, UK, 2017)

[14] Weinman, J.J., Zachary, B., Dugan, K., *et al.*: 'Toward integrated scene text reading', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (2), pp. 375–387

[15] Metwally, A.H., Khalil, M.I., Abbas, H.M.: 'Offline Arabic handwriting recognition using hidden Markov models and post-recognition lexicon matching'. 2017 12th Int. Conf. on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2017, pp. 238–243

[16] Rabi, M., Amrouch, M., Mahani, Z.: 'Recognition of cursive Arabic handwritten text using embedded training based on hidden Markov models', *Int. J. Pattern Recognit. Artif. Intell.*, 2018, **32**, (1), p. 1860007

[17] Rabi, M., Amrouch, M., Mahani, Z.: 'Cursive Arabic handwriting recognition system without explicit segmentation based on hidden Markov models', *J. Data Mining Digital Humanities*, 2018, pp. 1–7

[18] Zhou, P., Li, L., Tan, C.L.: 'Character recognition under severe perspective distortion'. 10th Int. Conf. on Document Analysis and Recognition, 2009. ICDAR'09, Barcelona, Spain, 2009, pp. 676–680

[19] Caner, G., Haritaoglu, I.: 'Shape-dna: effective character restoration and enhancement for Arabic text documents'. 2010 20th Int. Conf. on Pattern Recognition (ICPR), Istanbul, Turkey, 2010, pp. 2053–2056

[20] Kompalli, P.S.: 'Image document processing in a client-server system including privacy-preserving text recognition'. US Patent 9,847,974, 19 December 2017

[21] Märgner, V., Pal, U., Antonacopoulos, A., *et al.*: 'Document analysis and text recognition, 2018

[22] Wang, T., Wu, D.J., Coates, A., *et al.*: 'End-to-end text recognition with convolutional neural networks'. 2012 21st Int. Conf. on Pattern Recognition (ICPR), Tsukuba, Japan, 2012, pp. 3304–3308

[23] Sudholt, S., Fink, G.A.: 'Phocnet: a deep convolutional neural network for word spotting in handwritten documents'. 2016 15th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, People's Republic of China, 2016, pp. 277–282

[24] Liu, C.-L., Fink, G.A., Govindaraju, V., *et al.*: 'Special issue on deep learning for document analysis and recognition', 2018

[25] Lee, C.-Y., Bhardwaj, A., Di, W., *et al.*: 'Region-based discriminative feature pooling for scene text recognition'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 4050–4057

[26] Tian, S., Bhattacharya, U., Lu, S., *et al.*: 'Multilingual scene character recognition with co-occurrence of histogram of oriented gradients', *Pattern Recognit.*, 2016, **51**, pp. 125–134

[27] Yu, B., Wan, H.: 'Chinese text detection and recognition in natural scene using HOG and SVM', *DEStech Trans.Comput. Sci. Eng.*, 2016, pp. 148–152

[28] Yao, C., Bai, X., Shi, B., *et al.*: 'Strokelets: a learned multi-scale representation for scene text recognition'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 4042–4049

[29] Bai, X., Yao, C., Liu, W.: 'Strokelets: a learned multi-scale mid-level representation for scene text recognition', *IEEE Trans. Image Process.*, 2016, **25**, (6), pp. 2789–2802

[30] Yi, C., Tian, Y.: 'Scene text recognition in mobile applications by character descriptor and structure configuration', *IEEE Trans. Image Process.*, 2014, **23**, (7), pp. 2972–2982

[31] Sriman, B., Schomaker, L.: 'Object attention patches for text detection and recognition in scene images using sift'. ICPRAM (1), Lisbon, Portugal, 2015, pp. 304–311

[32] Jaderberg, M., Simonyan, K., Vedaldi, A., *et al.*: 'Synthetic data and artificial neural networks for natural scene text recognition', arXiv preprint arXiv:1406.2227

[33] Kumar, S., Kumar, K., Mishra, R.K., *et al.*: 'Scene text recognition using artificial neural network: a survey', *Int.J. Comput. Appl.*, 2016, **137**, (6), pp. 40–50

[34] Zhu, S., Zanibbi, R.: 'A text detection system for natural scenes with convolutional feature learning and cascaded classification'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 625–632

[35] Lu, S., Chen, T., Tian, S., *et al.*: 'Scene text extraction based on edges and support vector regression', *Int. J. Document Anal. Recogn. (IJDAR)*, 2015, **18**, (2), pp. 125–135

[36] Neumann, L.: 'Scene text localization and recognition in images and videos'. Ph.D. thesis, Department of Cybernetics Faculty of Electrical Engineering, Czech Technical University, 2017

[37] Hechri, A., Hmida, R., Mtibaa, A.: 'Robust road lanes and traffic signs recognition for driver assistance system', *Int. J. Comput. Sci. Eng.*, 2015, **10**, (1–2), pp. 202–209

[38] Ellahyani, A., El Ansari, M., El Jaafari, I.: 'Traffic sign detection and recognition based on random forests', *Appl. Soft Comput.*, 2016, **46**, pp. 805–815

[39] Salhi, A., Minaoui, B., Fakir, M., *et al.*: 'Traffic signs recognition using HP and HOG descriptors combined to MLP and SVM classifiers', *Traffic*, 2017, **8**, (11), pp. 526–530

[40] Saranya, K.C., Singhal, V.: 'Real-time prototype of driver assistance system for Indian road signs'. Int. Proc. on Advances in Soft Computing, Intelligent Systems and Applications, Singapore, Singapore, 2018, pp. 147–155

[41] Lai, Y., Wang, N., Yang, Y., *et al.*: 'Traffic signs recognition and classification based on deep feature learning'. 7th Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM), Madeira, Portugal, 2018, pp. 622–629

[42] Wang, K., Belongie, S.: 'Word spotting in the wild'. European Conf. on Computer Vision, Crete, Greece, 2010, pp. 591–604

[43] Goel, V., Mishra, A., Alahari, K., *et al.*: 'Whole is greater than sum of parts: recognizing scene text words'. 2013 12th Int. Conf. on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 2013, pp. 398–402

[44] Jaderberg, M., Vedaldi, A., Zisserman, A.: 'Deep features for text spotting'. European Conf. on Computer Vision, Zürich, Switzerland, 2014, pp. 512–528

[45] Shi, B., Wang, X., Lyu, P., *et al.*: 'Robust scene text recognition with automatic rectification'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 4168–4176

[46] 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', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, **39**, (11), pp. 2298–2304

[47] Yang, H., Li, S., Yin, X., *et al.*: 'Recurrent highway networks with attention mechanism for scene text recognition'. 2017 Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, 2017, pp. 1–8

[48] Bušta, M., Neumann, L., Matas, J.: 'Deep textspotter: an end-to-end trainable scene text localization and recognition framework'. 2017 IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2223–2231

[49] Lei, Z., Zhao, S., Song, H., *et al.*: 'Scene text recognition using residual convolutional recurrent neural network', *Mach. Vis. Appl.*, 2018, **29**, pp. 1–11

[50] Liu, Z., Li, Y., Ren, F., *et al.*: 'Squeezedtext: a real-time scene text recognition by binary convolutional encoder-decoder network'. Thirty-Second AAAI Conf. on Artificial Intelligence, New Orleans, LA, USA, 2018, pp. 23–28

[51] Liu, W., Chen, C., Wong, K.-Y.K.: 'Char-net: a character-aware neural network for distorted scene text recognition'. Thirty-Second AAAI Conf. on Artificial Intelligence, New Orleans, LA, USA, 2018, pp. 23–29

[52] Liao, M., Zhang, J., Wan, Z., *et al.*: 'Scene text recognition from two-dimensional perspective', arXiv preprint arXiv:1809.06508

[53] Gao, Y., Huang, Z., Dai, Y.: 'Double supervised network with attention mechanism for scene text recognition', arXiv preprint arXiv:1808.00677

[54] Lee, C.W., Jung, K., Kim, H.J.: 'Automatic text detection and removal in video sequences', *Pattern Recognit. Lett.*, 2003, **24**, (15), pp. 2607–2623

[55] Kim, D., Sohn, K.: 'Static text region detection in video sequences using color and orientation consistencies'. 19th Int. Conf. on Pattern Recognition, 2008. ICPR 2008, Tampa, FL, USA, 2008, pp. 1–4

[56] Phan, T.Q., Shivakumara, P., Lu, T., *et al.*: 'Recognition of video text through temporal integration'. 2013 12th Int. Conf. on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 2013, pp. 589–593

[57] Lee, S., Kim, J.: 'Complementary combination of holistic and component analysis for recognition of low-resolution video character images', *Pattern Recognit. Lett.*, 2008, **29**, (4), pp. 383–391

[58] Tang, X., Gao, X., Liu, J., *et al.*: 'A spatial-temporal approach for video caption detection and recognition', *IEEE Trans. Neural Netw.*, 2002, **13**, (4), pp. 961–971

[59] Elagouni, K., Garcia, C., Sébillot, P.: 'A comprehensive neural-based approach for text recognition in videos using natural language processing'.

Proc. of the 1st ACM Int. Conf. on Multimedia Retrieval, Trento, Italy, 2011, p. 23

[60] Shivakumara, P., Phan, T.Q., Lu, S.*, et al.*: 'Video character recognition through hierarchical classification'. 2011 Int. Conf. on Document Analysis and Recognition, Peking, People's Republic of China, 2011, pp. 131–135

[61] Khatri, M.J., Shetty, A., Gupta, A.*, et al.*: 'Video OCR for indexing and retrieval', *Int. J.Comput. Appl.*, 2015, **118**, (2), pp. 30–33

[62] Kulkarni, P., Bhagyashri, P., Joglekar, B.: 'An effective content based video analysis and retrieval using pattern indexing techniques'. 2015 Int. Conf. on Industrial Instrumentation and Control (ICIC), Pune, India, 2015

[63] Tikle, A.N., Vaidya, C., Dahiwale, P.: 'A survey of indexing techniques for large scale content-based image retrieval'. 2015 Int. Conf. on Electrical, Electronics, Signals, Communication and Optimization (EESCO), Andhra Pradesh, India, 2015

[64] Zayene, O., Touj, S.M., Hennebert, J.*, et al.*: 'Multi-dimensional long short-term memory networks for artificial Arabic text recognition in news video', *IET Comput. Vis.*, 2018, **12**, (5), pp. 710–719

[65] Zayene, O., Hennebert, J., Touj, S.M.*, et al.*: 'A dataset for Arabic text detection, tracking and recognition in news videos – activ'. 2015 13th Int. Conf. on Document Analysis and Recognition (ICDAR), Nancy, France, 2015

[66] Yousfi, S., Berrani, S.-A., Garcia, C.: 'Alif: a dataset for Arabic embedded text recognition in tv broadcast'. 2015 13th Int. Conf. on Document Analysis and Recognition (ICDAR), Nancy, France, 2015, pp. 1221–1225

[67] Yousfi, S., Berrani, S.-A., Garcia, C.: 'Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos'. 2015 13th Int. Conf. on Document Analysis and Recognition (ICDAR), Nancy, France, 2015, pp. 1026–1030

[68] Jain, M., Mathew, M., Jawahar, C.: 'Unconstrained scene text and video text recognition for Arabic script'. 2017 1st Int. Workshop on Arabic Script Analysis and Recognition (ASAR), Lorraine, France, 2017, pp. 26–30

[69] Bhunia, A.K., Kumar, G., Roy, P.P.*, et al.*: 'Text recognition in scene image and video frame using color channel selection', *Multimedia Tools Appl.*, 2018, **77**, (7), pp. 8551–8578

[70] Lu, W., Sun, H., Chu, J.*, et al.*: 'A novel approach for video text detection and recognition based on a corner response feature map and transferred deep convolutional neural network', *IEEE Access*, 2018, **6**, pp. 40198–40211

[71] Yan, X., Siyuan, S., Ziming, Q.*, et al.*: 'End-to-end subtitle detection and recognition for videos in east Asian languages via CNN ensemble', *Signal Process., Image Commun.*, 2018, **60**, pp. 131–143

[72] Naz, S., Umar, A.I., Ahmad, R.*, et al.*: 'Offline cursive urdu-nastaliq script recognition using multidimensional recurrent neural networks', *Neurocomputing*, 2016, **177**, pp. 228–241

[73] Javed, N., Shabbir, S., Siddiqi, I.*, et al.*: 'Classification of Urdu ligatures using convolutional neural networks-a novel approach'. 2017 Int. Conf. on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2017, pp. 93–97

[74] Naz, S., Umar, A.I., Ahmad, R.*, et al.*: 'Urdu nastaliq recognition using convolutional–recursive deep learning', *Neurocomputing*, 2017, **243**, pp. 80–87

[75] Naseer, A., Zafar, K.: 'Comparative analysis of raw images and meta feature based Urdu OCR using CNN and lstm', *Int. J. Adv. Comput. Sci. Appl.*, 2018, **9**, (1), pp. 419–424

[76] Uddin, I., Siddiqi, I., Khalid, S.*, et al.*: 'Segmentation-free optical character recognition for printed urdu text', *EURASIP J. Image Video Process.*, 2017, **2017**, (1), p. 62

[77] Shamsher, I., Ahmad, Z., Orakzai, J.K.*, et al.*: 'OCR for printed Urdu script using feed forward neural network'. Proc. of World Academy of Science, Engineering and Technology, vol. **23**, Dubai, UAE, 2007

[78] Tariq, J., Nauman, U., Naru, M.U.: 'Softconverter: a novel approach to construct OCR for printed Urdu isolated characters'. 2010 2nd Int. Conf. on Computer Engineering and Technology (ICCET), Chengdu, People's Republic of China, 2010, vol. 3, p. V3V495V3–495

[79] Sardar, S., Wahab, A.: 'Optical character recognition system for urdu'. Int. Conf. on Information and Emerging Technologies, Karachi, Pakistan, 2010

[80] Ahmad, I., Mahmoud, S.A., Fink, G.A.: 'Open-vocabulary recognition of machine-printed Arabic text using hidden Markov models', *Pattern Recognit.*, 2016, **51**, pp. 97–111

[81] Khemiri, A., Echi, A.K., Belaid, A.*, et al.*: 'Arabic handwritten words off-line recognition based on hmms and dbns'. 2015 13th Int. Conf. on Document Analysis and Recognition (ICDAR), Nancy, France, 2015, pp. 51–55

[82] Javed, S.T., Hussain, S.: 'Segmentation based Urdu nastalique OCR'. Iberoamerican Congress on Pattern Recognition, Havana, Cuba, 2013, pp. 41–49

[83] Aved, S.T., Hussain, S., Maqbool, A.*, et al.*: 'Segmentation free nastalique Urdu OCR', *World Academy of Sci., Eng. Technol.*, 2010, **46**, pp. 456–461

[84] Lehal, G.S.: 'Choice of recognizable units for Urdu OCR'. Proceeding of the workshop on document analysis and recognition, New York, NY, USA, 2012, pp. 79–85

[85] Sagheer, M.W., Nobile, N., He, C.L.*, et al.*: 'A novel handwritten Urdu word spotting based on connected components analysis'. 2010 20th Int. Conf. on Pattern Recognition (ICPR), Istanbul, Turkey, 2010, pp. 2013–2016

[86] Abidi, A., Jamil, A., Siddiqi, I.*, et al.*: 'Word spotting based retrieval of Urdu handwritten documents'. 2012 Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), Bari, Italy, 2012, pp. 331–336

[87] Ul-Hasan, A., Ahmed, S.B., Rashid, F.*, et al.*: 'Offline printed Urdu nastaleeq script recognition with bidirectional LSTM networks'. 2013 12th Int. Conf. on Document Analysis and Recognition, Washington, DC, USA, 2013, pp. 1061–1065

[88] Naz, S., Umar, A.I., Ahmed, R.*, et al.*: 'Urdu nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks', *SpringerPlus*, 2016, **5**, (1), p. 2010

[89] Sabbour, N., Shafait, F.: 'A segmentation-free approach to Arabic and Urdu OCR'. IS&T/SPIE Electronic Imaging, Int. Society for Optics and Photonics, San Francisco, CA, USA, 2013, pp. 86580N–86580N

[90] Center for language engineering. http://http://www.cle.org.pk/, accessed: 2019-04-15

[91] Sagheer, M.W., He, C.L., Nobile, N.*, et al.*: 'Holistic Urdu handwritten word recognition using support vector machine'. 2010 20th Int. Conf. on Pattern Recognition (ICPR), Istanbul, Turkey, 2010, pp. 1900–1903

[92] Raza, A., Siddiqi, I., Abidi, A.*, et al.*: 'An unconstrained benchmark Urdu handwritten sentence database with automatic line segmentation'. 2012 Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), Bari, Italy, 2012, pp. 491–496

[93] Hayat, U., Aatif, M., Zeeshan, O.*, et al.*: 'Ligature recognition in Urdu caption text using deep convolutional neural networks'. 2018 14th Int. Conf. on Emerging Technologies (ICET), Islamabad, Pakistan, 2018, pp. 1–6

[94] Tayyab, B.U., Naeem, M.F., Ul-Hasan, A.*, et al.*: 'A multi-faceted OCR framework for artificial Urdu news ticker text recognition'. 2018 13th IAPR Int. Workshop on Document Analysis Systems (DAS), Vienna, Austria, 2018, pp. 211–216

[95] Mirza, A., Siddiqi, I., Mustufa, S.G.*, et al.*: 'Impact of pre-processing on recognition of cursive video text'. 9th Iberian Conf. on Pattern Recognition and Image Analysis (IbPRIA 2019), Madrid, Spain, 2019

[96] Otsu, N.: 'A threshold selection method from gray-level histograms', *IEEE Trans. Syst. Man Cybernet.*, 1979, **9**, (1), pp. 62–66

[97] Niblack, W.: '*An introduction to digital image processing*', vol. **34** (Prentice-Hall, Englewood Cliffs, 1986)

[98] Sauvola, J., Pietikäinen, M.: 'Adaptive document image binarization', *Pattern Recognit.*, 2000, **33**, (2), pp. 225–236

[99] Feng, M.-L., Tan, Y.-P.: 'Contrast adaptive binarization of low quality document images', *IEICE Electron. Express*, 2004, **1**, (16), pp. 501–506

[100] Wolf, C., Jolion, J.-M.: 'Extraction and recognition of artificial text in multimedia documents', *Formal Pattern Anal. Appl.*, 2004, **6**, (4), pp. 309–326

[101] Bodapati, J.D., Suvarna, B., N, V.: 'Role of deep neural features vs hand crafted features for hand written digit recognition', *Int. J. Recent Technol. Eng. (IJRTE)*, 2019, **7**, pp. 147–152

[102] Nanni, L., Ghidoni, S., Brahnam, S.: 'Handcrafted vs. non-handcrafted features for computer vision classification', *Pattern Recognit.*, 2017, **71**, pp. 158–172

[103] Alshazly, H., Linse, C., Barth, E.*, et al.*: 'Handcrafted versus CNN features for ear recognition', *Symmetry*, 2019, **11**, (12), p. 1493

[104] Ahmad, Z., Orakzai, J.K., Shamsher, I.*, et al.*: 'Urdu nastaleeq optical character recognition', *Int. J. Comput. Inf. Eng.*, 2007, **1**, (8), pp. 2380–2383

[105] Akram, Q., Hussain, S., Adeeba, F.*, et al.*: 'Framework of Urdu nastalique optical character recognition system'. Proc. of Conf. on Language and Technology (CLT), Karachi, Pakistan, 2014, pp. 1–7

[106] Hussain, S., Ali, S., Akram, Q.U.A.: 'Nastalique segmentation-based approach for Urdu OCR', *Int. J. Doc. Anal. Recognit. (IJDAR)*, 2015, **18**, (4), pp. 357–374

[107] Ahmed, S.B., Naz, S., Razzak, M.I.*, et al.*: 'Evaluation of cursive and non-cursive scripts using recurrent neural networks', *Neural Comput. Appl.*, 2016, **27**, (3), pp. 603–613

[108] Naz, S., Umar, A.I., Ahmad, R.*, et al.*: 'Urdu nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features', *Neural Comput. Appl.*, 2017, **28**, (2), pp. 219–231