Mirza *et al. EURASIP Journal on Image and Video Processing* (2020) 2020:34
https://doi.org/10.1186/s13640-020-00523-5

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Detection and recognition of cursive text from video frames

Ali Mirza[1*], Ossama Zeshan[1], Muhammad Atif[1] and Imran Siddiqi[1]

Abstract

Textual content appearing in videos represents an interesting index for semantic retrieval of videos (from archives), generation of alerts (live streams), as well as high level applications like opinion mining and content summarization. The key components of such systems require detection and recognition of textual content which also make the subject of our study. This paper presents a comprehensive framework for detection and recognition of textual content in video frames. More specifically, we target cursive scripts taking Urdu text as a case study. Detection of textual regions in video frames is carried out by fine-tuning deep neural networks based object detectors for the specific case of text detection. Script of the detected textual content is identified using convoluational neural networks (CNNs), while for recognition, we propose a UrduNet, a combination of CNNs and long short- term memory (LSTM) networks. A benchmark dataset containing cursive text with more than 13,000 video frame is also developed. A comprehensive series of experiments is carried out reporting an F-measure of 88.3% for detection while a recognition rate of 87%.

Keywords: Text detection, Text recognition, Script identification, Deep neural networks (DNNs), Convolutional neural networks (CNNs), Long short-term memory (LSTM) networks, Caption text, Cursive text

1 Introduction

In the recent years, there has been a tremendous increase in the amount of digital multimedia data, especially the video content, both in the form of video archives and live streams. According to statistics [1], 300 h of video is being uploaded every minute on the YouTube. A key factor responsible for this enormous increase is the availability of low-cost smart phones equipped with cameras. With such huge collections of data, there is a need to have efficient as well as effective retrieval techniques allowing users retrieve the desired content. Traditionally, videos are mostly stored with user assigned annotations or keywords which are called tags. When a content is to be searched, a keyword provided as query is matched with these tags to retrieve the relevant content. The assigned tags, naturally, cannot encompass the rich video content leading to a constrained retrieval. A better and more effective strategy is to search within the actual content rather than simply matching the tags, i.e., content based (image or) video retrieval. CBVR systems have been researched and developed for a long time and allow a smarter way of retrieving the desired content. The term content may refer to the visual content (for example, objects or persons in the video), audio content (the spoken keywords for instance), or the textual content (News tickers, anchor names, score cards, etc.). Among these, the focus of our current study lies on textual content. More specifically, we target a smart retrieval system that exploits the textual content in videos as an index.

The textual content in video can be categorized into two broad classes, scene text, and caption text. Scene text (Fig. 1) is captured through camera during the video recording process and may not always be correlated with the content. Examples of scene text include advertisement banners, sign boards, and text on a T-shirt. Scene text is commonly employed for applications like robot navigation and assistance systems for the visually impaired. Artificial or caption text (Fig. 2), on the other hand, is superimposed on the video and typical examples include

*Correspondence: alimirza@bahria.edu.pk
[1]Bahria University, Islamabad, Pakistan

Fig. 1 Examples of scene text

News tickers, movie credits, and score cards. Caption text is generally correlated with the video content and is mostly applied for semantic retrieval of videos. The focus of our present study also lies on the caption text.

The key components of a textual content based indexing and retrieval system include detection of text regions, extraction of text (segmentation from background), identification of script (for multi-script videos), and finally recognition of text (video OCR). Detection of text can be carried out using unsupervised [2–4], supervised [5–7], or hybrid [8, 9] approaches. Unsupervised text detection relies on image analysis techniques to discriminate between text and non-text regions. Supervised methods, on the other hand, involve training a learning algorithm with examples of text and non-text regions to discriminate between the two. In some cases, a combination of the two techniques is also employed where the candidate text regions identified by unsupervised methods are validated through a supervised approach.

Once the text is detected, it can be fed to the recognition engine. In case of videos which include text in multiple scripts, an additional module to identify the script is required so that each type of text can be recognized through its respective OCR. While mature recognition systems are available of text in the Roman script (ABBYY Fine Reader and Tesseract [10], etc.), recognition of cursive scripts remains a challenging task. Furthermore, as opposed to document images which are scanned at high resolution, video text is mostly in low resolution (Fig. 3) and can appear on complex backgrounds making its recognition more challenging.

This paper presents a comprehensive framework for video text detection and recognition in a multi-script environment. The key highlights of this study are outlined in the following.

- Development of a comprehensive dataset of video frames with ground truth information supporting evaluation of detection and recognition tasks.
- Investigation of state-of-the-art deep learning based object detectors, fine-tuning them to detection of textual content.
- Video text script identification using convolutional neural networks (CNN).
- Recognition of cursive (Urdu) video text using a combination of CNN and LSTM (UrduNet).
- Validation of proposed techniques using a comprehensive series of experiments.

This paper is organized as follows. In the next section, we present an overview of the current state-of-the-art from the view point of text detection, script identification, and text recognition. In Section 3, we present the database developed in our study along with the ground truth information. Section 4 presents the details of the proposed framework while Section 5 details the experimental protocol, the realized results, and the corresponding discussion. Finally, Section 6 concludes the paper with a discussion on open challenges on this subject.

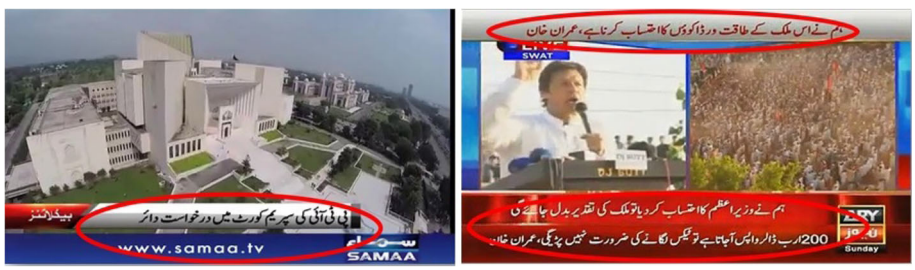

Fig. 2 Examples of caption text

Fig. 3 Low-resolution text in video frames

2 Background

Detection and recognition of textual content in videos, images, documents, and natural scenes has been investigated for more than four decades. The domain has matured progressively over the years starting with trivial systems recognizing isolated digits and characters to complex end-to-end systems capable of reading text in natural scenes. This section presents an overview of the notable contributions to detection and recognition of textual content in images and videos. Comprehensive and detailed surveys on the problem can found in [11–15]. We organize our discussion in three main sections. We first discuss the text detection problem followed by an overview of script identification techniques. At the end, we present the current state-of-the-art and open challenges from the view point of text recognition.

2.1 Text detection

Text detection refers to localization of textual content in images. Techniques proposed for detection of text are typically categorized into unsupervised and supervised approaches. While unsupervised methods primarily rely on image analysis techniques to segment text from background, supervised techniques involve training a learning algorithm to discriminate between text and non-text regions.

Unsupervised text detection techniques include edge-based methods [2, 3, 16] which (assume and) exploit the high contrast between text and its background; connected component-based methods [17, 18] which mostly rely on the color/intensity of text pixels and texture-based methods [19, 20] which consider textual content in the image as a unique texture that distinguishes itself from the non-text regions. Texture based methods have remained a popular choice of researchers and features based on Gabor filters [21], wavelets [22], curvelets [23], local binary patterns (LBP) [24], discrete cosine transformation (DCT) [25], histograms of oriented gradients (HoG) [26], and Fourier transformation [27] have been investigated in the literature. Another common category of techniques includes color-based methods [28, 29] which are similar in many aspects to the component-based methods and employ color information of text pixels to distinguish it from non-text regions.

Supervised approaches for detection of textual content typically employ state-of-the-art learning algorithms which are trained on examples of text and non-text blocks either using pixel values or by first extracting relevant features. Classifiers like Naïve Bayes [30], support vector machine (SVM) [31], artificial neural network (ANN) [8, 32], and deep neural networks (DNN) [33] have been investigated over the years.

In the recent years, deep learning based solutions have been widely applied to a variety of classification problems and have outperformed the traditional techniques. A major development contributing to the current fame of deep learning was the application of convolutional neural networks (CNNs) by Krizhevsky et al. [34] on the ImageNet Large Scale Visual Recognition competition [35], which greatly reduced the error rates. Since then, CNNs are considered to be state-of-the-art feature extractors and classifiers [36, 37] and have been applied to a number of recognition tasks. While traditional CNNs are typically employed for object classification, region-based convolutional networks (R-CNN) [38] and their further enhancements Fast R-CNN [39] and Faster R-CNN [40] represent common object detectors. In addition to different variants of R-CNN, a number of new architectures have also been proposed in the recent years for real time object detection. The most notable of these include YOLO (You Only Look Once) [41] and SSD (Single Shot Detector) [42]. These object detectors can be fine-tuned with textual data to serve as text detectors and are likely to provide good results.

Among deep learning based techniques adapted for text detection, Huang et al. [43] employed sliding windows with CNNs to detect textual regions in low-resolution scene images. Likewise, fully convolutional networks are explored for detection of textual regions in [44] and the technique is evaluated on various ICDAR datasets. A similar work is presented by Gupta et al. [45] where CNNs are trained using synthetic data for detection of text at multiple scales from natural images. Another method called "SegLink," is proposed in [46] that relies on decomposing the text into segments (oriented boxes of words or lines) and links (connecting two adjacent segments). The segments and links are detected using fully convolutional networks at multiple scales and combined together to detect the complete text line. In [47], vertical anchor-based method is reported that predicts text and non-text scores of fixed size regions and reports high detection performance on the ICDAR 2013 and ICDAR 2015 datasets. In another notable work, Wang et al. [48] present a framework based on conditional random field (CRF) to detect text in scene images. Authors define a cost function by considering the color, stroke, shape, and spatial features with CNN for effective detection of textual regions.

Among other end-to-end trainable deep neural networks based systems, Liao et al. [7] present a system called "TextBoxes" which detects text in natural images in a single forward pass network. The technique was later extended to "TextBoxes++" and was evaluated on four public databases outperforming the state-of-the-art. He et al. [6] improved the convolutional layer of CNNs to detect text with arbitrary orientation. EAST [49] is another well-known scene text detector that provides promising results in challenging scenarios. In another study [5], an ensemble of CNNs is trained on synthetic data to detect video text in East Asian languages.

Summarizing, it can be concluded that the problem of text detection has been mostly dominated by the application of different deep learning architectures in the recent years. The availability of benchmark datasets has also contributed to the rapid developments in this area. Among open problems, development of a generic text detector that could work in multi-script environment remains a challenging issue.

In the next section, we discuss different techniques for recognition of script of the text from documents and video images.

2.2 Script recognition

In many cases, videos may contain textual content in more than one script. These scripts must be identified prior to feeding the text regions to the respective OCR engines for recognition. Script recognition has been studied by researchers for text in video images as well as printed and handwritten documents [50, 51]. Recognition of script in video text is naturally much more challenging as opposed to printed or handwritten documents due to low resolution of text and in some cases complex backgrounds [52, 53]. From simple methods based on template matching [54] to sophisticated structural [55] and statistical [56] features, a number of techniques have been reported in the literature. Among various features exploited to characterize the script, texture-based features [53, 57–59] are known to be very effective reporting high classification rates. Textural measures like Gabor filters [60], LBP [24], and gray level co-occurrence matrix (GLCM) [61] have been investigated in a number of studies. The extracted features are typically fed to traditional classifiers to discriminate between the script classes under study. Among well-known methods, Zhao et al. [62] employed "Spatial Gradient Features (SGF)" to characterize script of text using a dataset of six different scripts. A similar study with "Gradient Angular Feature (GAF)" is reported in [63] where authors presented "Potential Text Candidate (PTC)" method for studying the cursiveness of text with histogram operations. Sharma et al. [64] proposed script identification using "Gradient Local Auto-Correlation (GLAC)" for English, Bengali, and Hindi script in low-resolution video frames. A recent comprehensive survey on script identification can be found in [50]. A competition on this problem was also organized in conjunction with ICDAR 2015 [65]. The competition involved four challenging tasks with 10 different languages, and among the submitted systems, Google Inc. was declared the winner of the competition.

Among recent contributions to script identification, Jieru et al. [66] combine a CNN and RNN into a single end-to-end trainable network. The technique is evaluated on multiple datasets and reports high identification rates. In [67], authors propose a set of mid-level features for script identification with very less labeled data. Experiments on CVSI dataset report an identification rate of more than 96%. Gomez et al. [68] employed Naïve Bayes classifier with convolutional features to identify script in unconstrained scene text. The work was later extended to apply patch-based classification using CNNs [69]. In other recent works, transfer learning and fine-tuning with AlexNet and VGG-16 are explored in [70] for script identification. Bag of visual words model is investigated in [71] by using convolutional features extracted from image patches in the form of triplets. Bhunia et al. [72] propose a CNN-LSTM framework to extract local as well as global features. The features are weighted dynamically, and the technique is evaluated on four public datasets reporting high identification rates.

Summarizing, like many other problems, deep learning-based methods have been the dominant technique for script identification in the recent years. While the initial research primarily focused on document images, script

identification of text appearing in videos has been an attractive research theme for many years now. Though many sophisticated systems have come to scene, low resolution of video images and the high similarity between different scripts are keeping it an active research area.

In the next section, we discuss different techniques for text recognition from documents and video images.

2.3 Text recognition

Recognition of text, generally termed as OCR (Optical Character Recognition) is one of the most classical pattern recognition tasks that has been explored for decades. Recognition systems have been investigated for printed as well handwritten documents (scanned or camera-based), text in natural scene images, and the caption text appearing in videos. State-of-the-art recognition systems (for instance Google Tesseract [10], Abbyy FineReader, etc.) are known to report near to 100% recognition rates for textual content in a number of scripts. Recognition of text in cursive scripts, however, still remains challenging, especially for the text appearing in videos.

From the view point of document OCRs, research endeavors can be categorized into two main classes, analytical (character-based) and holistic (word-based) techniques. Analytical techniques which work either on isolated characters or first segment the text into characters. For text recognition in document images, a number of techniques have been presented both at character (analytical) and word (holistic) levels. Typically, techniques like graph-based models [73–75], Bayesian classifiers [76, 77], and Hidden Markov Models [78–80], etc., have been explored for character level recognition of text. Likewise, a number of features and classifiers have been investigated for word level recognition [81–83]. A number of recent studies also employed deep learning-based solutions for analytical as well holistic recognition of text [84, 85].

Unlike document images, recognition of text from scene images offers a more challenging scenario due to different positions of camera while capturing the text, non-uniform illumination, and complex backgrounds. Among hand-engineered features, popular descriptors investigated for detection of natural scene text include the Scale Invariant Feature Transform (SIFT) [86, 87], Strokelets [88], and Histogram of Oriented Gradients (HOG) [89, 90], etc. Likewise, ANNs [91–93] and SVMs [94, 95] have been commonly employed as classifiers. In addition to recognition, techniques based on word spotting have also been investigated on scene text images [96, 97]. Recognition of text in road signs also represents an important subproblem within the umbrella of scene text recognition and has been explored in a number of studies [98–100].

A recent trend in text recognition has been the combination of convolutional and recurrent neural networks [101–105] where the CNN part serves to map the raw text images to effective feature representations while the recurrent part exploits the feature sequences to predict the transcription. In addition to the standard C-RNN, a number of enhancements have been proposed in the network architectures [106–109] to deal with the challenges of a scene text.

From the perspective of recognition of caption text, a key challenge, as discussed earlier, is the low resolution of text. A number of studies address this problem as a pre-processing step and combine the information from multiple frames to produce a high-resolution image which is subsequently fed to the recognizer [110, 111]. Recognition of caption text has been mostly employed for indexing and retrieval applications [112, 113].

In the context of cursive text, a holistic technique based on multi-dimensional LSTMs is presented in [114] for recognition of Arabic video text. The technique is evaluated on two datasets ACTiV [115] and the ALIF [116, 117] and reports high recognition rates. A similar work is reported in [118] where a combination of CNN and LSTM is employed to recognize Arabic text in video frames. Another deep learning-based solution is presented in [119] where Lu et al. compare the performance of different pre-trained ConvNets for detection and recognition of caption text. CNNs are also employed for recognition of Chinese video text in [5].

Recognition of Urdu text has recently received significant research attention. Most of the developed systems mainly target digitized printed documents. Similar to other cursive scripts, recognition techniques are categorized into analytical (segmentation-free) and holistic (segmentation-based) methods. Unlike other scripts however, segmentation of Urdu text into characters is a challenging problem [120]. As a result, a number of implicit segmentation-based techniques have been recently proposed where the learning algorithm is provided with the text line images and the corresponding output transcription to learn different shapes of a character as well as the segmentation points [121–123]. Likewise, in holistic approaches, the word boundaries are difficult to identify hence subwords (ligatures) are typically employed as recognition units in these methods.

Among notable holistic approaches, HMMs have been widely employed for recognition of ligatures [124–126]. These techniques typically employ sliding windows to extract features from ligature images which are projected in the quantized feature space, hence representing each ligature image as a sequence. In some cases, the main body and dots are separately recognized [127] to reduce the total number of unique classes which can be very high (Urdu, for example, has more than 26,000 unique ligatures [128]). The implicit segmentation-based recognition techniques mostly employ different variants of LSTMs [121, 129, 130] with a connectionist

temporal classification (CTC) output layer to recognize characters. A significant proportion of studies targeting recognition of Urdu text employ the publicly available UPTI [131] and CLE [132] datasets. The UPTI dataset comprises more than 10,000 synthetic text lines while the CLE dataset consists of scanned images of printed Urdu books as well as a collection of high frequency ligatures.

While recognition of printed Urdu text in document images has progressively matured in the recent years, research endeavors targeting caption text are fairly limited. A holistic approach for recognition of a small set of Urdu ligatures (collected from video text) is presented in [133]. Pre-trained ConvNets are employed for feature extraction and classification, and though high recognition rates are reported, the number of considered ligature classes is very limited (290 ligature classes). Likewise, an implicit segmentation-based technique using LSTMs is presented for recognition of Urdu News tickers in [134]. The experimental study is carried out on a private dataset of videos, and the recognizer performance is compared with a commercial OCR.

After having discussed the significant contributions to text detection, script identification, and text recognition, we now present the dataset that has been developed to support evaluation of text detection and recognition modules.

3 Dataset

For experimental study of our system, we have collected and labeled a comprehensive dataset of video frames. The frames are labeled to allow evaluation of text detection, script identification, and text recognition performance. For each frame, the location of text regions, the script information, and the ground truth transcription are stored.

The first step in database development is the collection of videos. We have collected 60 videos by recording live streams from five different News channels. All videos are recorded at a resolution of 900 × 600 and a frame rate of 25 fps. While the videos contain textual content in both Urdu and English, videos from four of the channels have dominant occurrences of Urdu text while those from one channel mostly contain textual content in English. Since successive frames in a video contain redundant information, we extract one frame every two seconds for labeling. The main reason of extracting a single frame every 2 s is to ensure that the collected frames have different textual content. This allows variation in training and test data as opposed to the case where a sequence of frames contains (mostly) similar textual content. Having as many unique words and character combinations allows the learning algorithm generalize better.

To facilitate the labeling process and standardize the ground truth data, a comprehensive labeling tool has been developed that allows storing the location of each textual region in a frame along with its ground truth transcription. A screen shot of the developed tool is presented in Fig. 4. The tool allows loading frames in a video and labeling them one by one for text locations as well as ground truth transcription. The ground truth information of each frame is stored as an XML file and comprises frame meta data (video and channel details) and information about text regions in the frame. For each script (English & Urdu

Fig. 4 Screenshot of ground truth labeling tool

```
<?xml version="1.0" encoding="utf-8"?>
<VideoLabel>
  <FrameMetaData>
    <Video>test_images</Video>
    <Channel>Samaa News</Channel>
    <FrameNo>image9</FrameNo>
  </FrameMetaData>
  <TextFeeds TotalFeeds="7">
    <UrduFeeds TotalUrduFeeds="3">
      <TextLine ID="1" TextType="Artificial" X="130" Y="285" Width="74" Height="31" Text="غریدہ فاروقی" />
      <TextLine ID="2" TextType="Artificial" X="733" Y="516" Width="72" Height="21" Text="نیوز" />
      <TextLine ID="3" TextType="Artificial" X="105" Y="517" Width="549" Height="51" Text="پیرس: ملک بھر میں 70 ہزار پولنگ اسٹیشنز قائم، خبرایجنسی" />
    </UrduFeeds>
    <EnglishFeeds TotalEnglishFeeds="4">
      <TextLine ID="1" TextType="Artificial" X="708" Y="487" Width="123" Height="25" Text="express" />
      <TextLine ID="2" TextType="Artificial" X="723" Y="547" Width="92" Height="20" Text="11:09 am" />
      <TextLine ID="3" TextType="Artificial" X="5" Y="470" Width="201" Height="33" Text="chi for 90 days" />
      <TextLine ID="4" TextType="Artificial" X="294" Y="468" Width="406" Height="33" Text="earthuake jolts hairpur and" />
    </EnglishFeeds>
  </TextFeeds>
</VideoLabel>
```

Fig. 5 Ground truth information of a labeled frame in an XML file

in our case), we store information on total number of text lines, and for each line, we store a unique ID, the type of text (scene text or artificial text), the location of text region within the frame, and the transcription of text. The ground truth information of an example frame (stored in the XML format) is illustrated in Fig. 5 while a summary of the labeled data in terms of number of videos, number of frames, and number of text lines is presented in Table 1.

4 Methods

This section presents the details of the proposed framework which is summarized in Fig. 6. The overall system comprises of three main modules, text detector, script identifier, and text recognizer. On top of these modules, a wide range of systems can be developed at the application layer including indexing and retrieval, key-word-based alert generation, and content summarization. The first module, text detector is responsible for identifying and localizing all textual content in a frame. Since text can be in more than one script (within the same frame), the detected textual regions are fed to the script identification module which separates the text lines as a function of the script (English and Urdu being the two scripts considered in the present study). The text is finally passed to the respective recognition engines of each script to convert the images of text lines into strings which can be subsequently employed for a number of applications. Each of these modules is discussed in detail in the following.

4.1 Text detection

The first step in the proposed framework is the detection of candidate text regions from the extracted video frames. For detection of textual content in a given frame, we have employed state-of-the-art convolutional neural networks (CNN) -based object detectors. Although, many object detectors are trained with thousands of class examples and provide high accuracy in detection and recognition of different objects, these object detectors cannot be directly applied to identify text regions in images. These models have to be tuned to the specific problem of discrimination of text from non-text regions. The convolutional base of these models can be trained from scratch or, known pre-trained models (VGG, Inception, or ResNet) can be fine-tuned by training them on text and non-text regions. In our study, we investigated the following object detectors for localization of text regions. The idea is to study which of these can be better adapted for text detection problem.

- Faster R-CNN
- Single Shot Detector (SSD)
- Efficient and Accurate Scene Text detection pipeline (EAST)

Faster R-CNN [40] is an enhanced version of its predecessors R-CNN [38] and Fast R-CNN [39]. Faster R-CNN merges a region proposal network with Fast R-CNN for effective and efficient localization and recognition of objects. The SSD (Single Shot multibox Detector) architecture was proposed by Liu et al. [42] and reported high precision on object detection on standard datasets like PascalVOC and COCO. The architecture has an input

Table 1 Channel videos and video image statistics

S#	Channel	Videos	Labeled frames	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	8861
4	Express News	10	2424	8536	6755
5	PTV World	10	2354	0	10,459
	Total	56	13,546	40,470	34,091

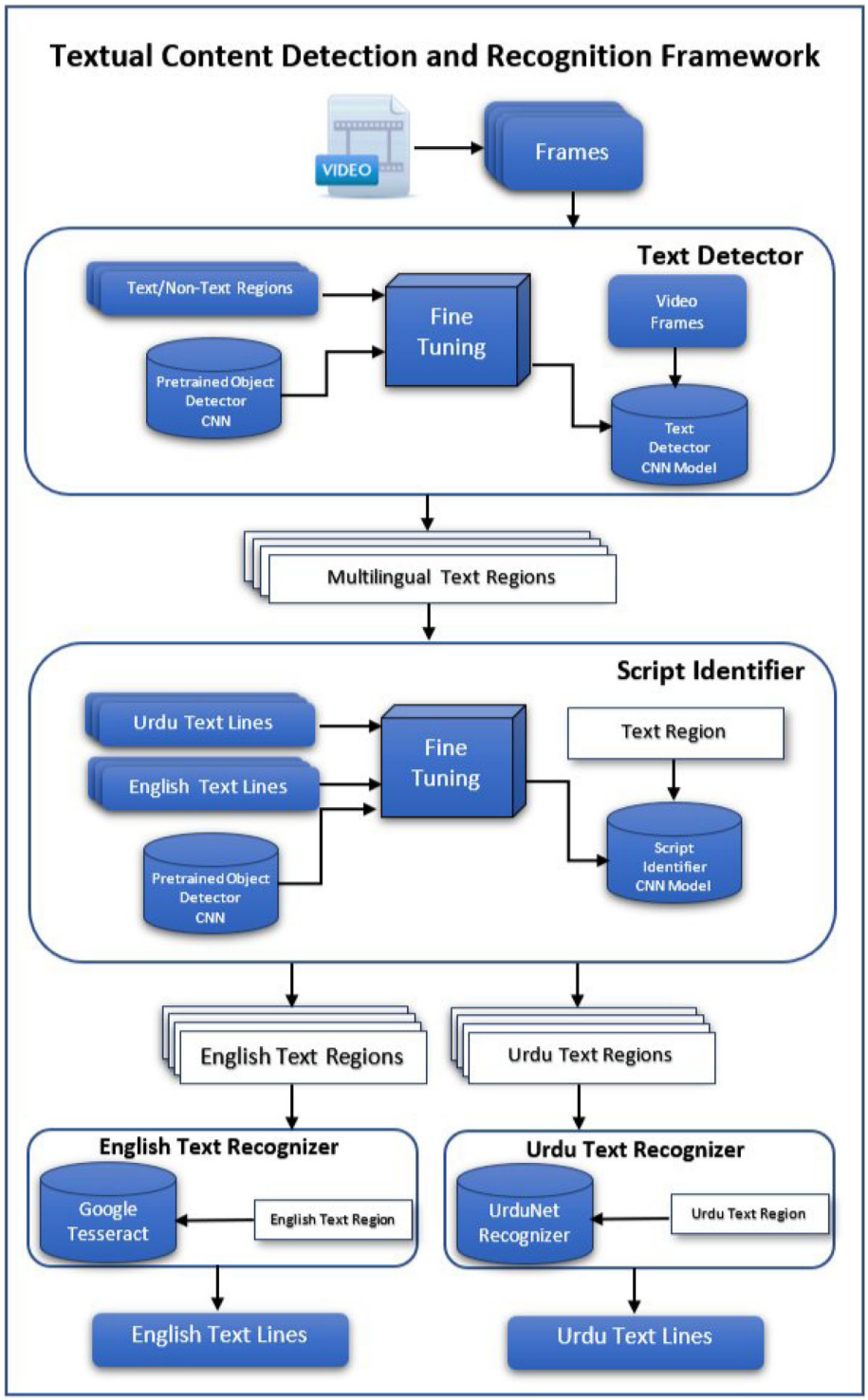

Fig. 6 An overview of proposed framework

size of 300 × 300 × 3 and builds on VGG-16 model discarding the fully connected layers. EAST (Efficient and Accurate Scene Text detection pipeline) [49] localizes text regions in natural scene images along with the geometries of text. It can predict complete text lines as well as single words. Prediction of geometrical shape of text is

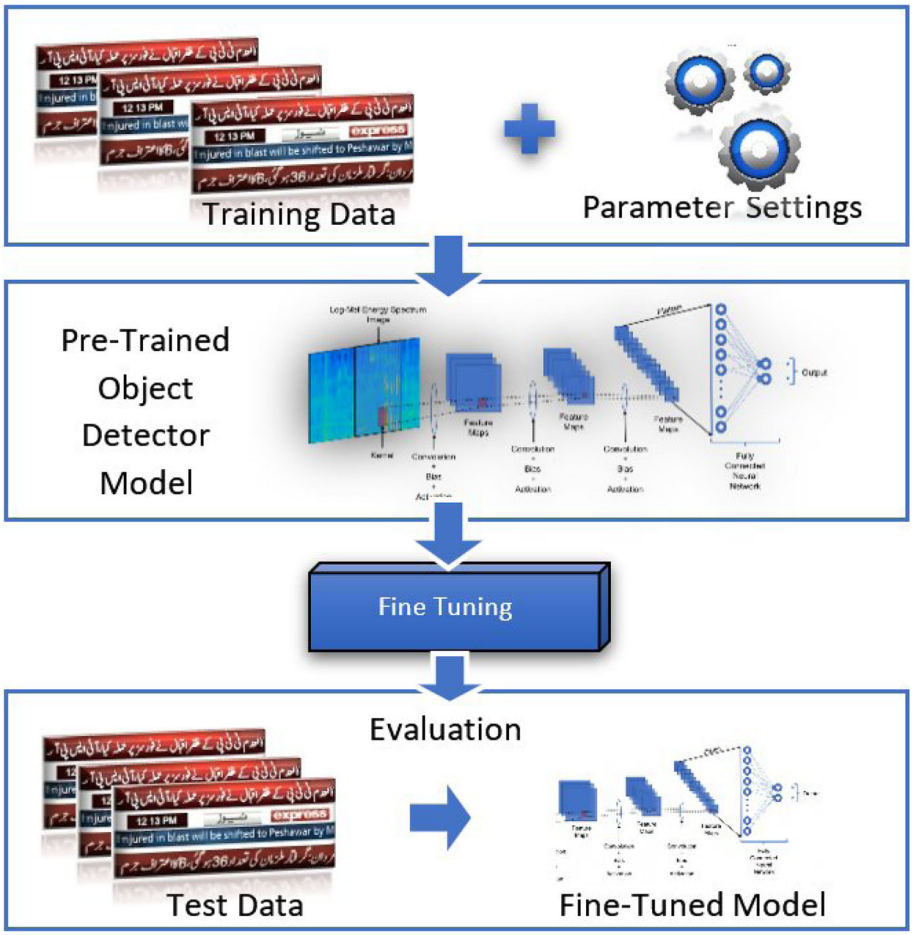

Fig. 7 Overview of fine-tuning object detectors for text

also a unique characteristic of this model. While Faster R-CNN and SSD are generalized object detectors, EAST was specifically designed to detect text from scene images. In our study however, it did not report acceptable detection performance once applied directly to the detection of caption text from video images. Consequently, in addition to Faster R-CNN and SSD, EAST was also fine-tuned to our dataset.

To investigate the performance of different object detectors on our specific problem, we carried out a comprehensive series of experiments by training these three models for various setting of hyper parameters. Text regions, containing Urdu and English text lines, are given as training examples to these detectors. Once tuned, the detection performance is evaluated using the test set of images. The overall flow of the text detection is summarized in Fig. 7.

Fig. 8 Text detector output on sample video frames

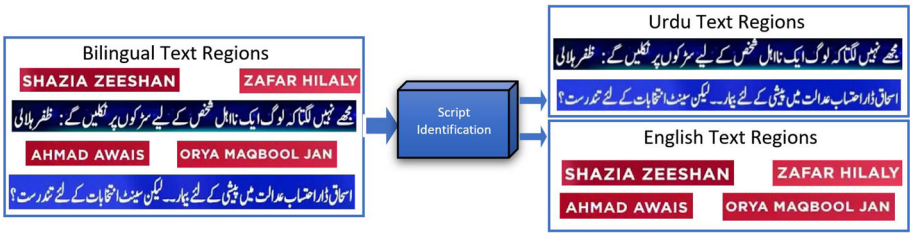

Fig. 9 Script Identification to separate text as a function of script

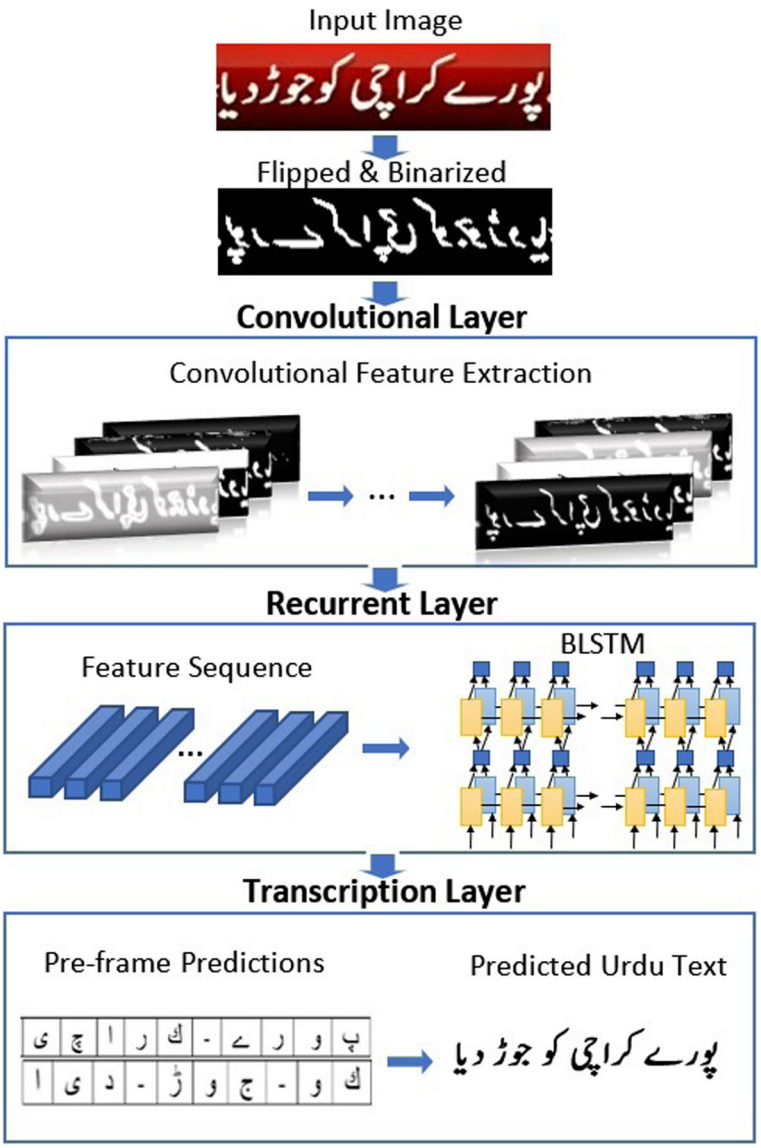

Fig. 10 UrudNet: overview of steps involved in recognition of text lines

Fig. 11 Sample images of text lines for model training

The final fine-tuned model takes a video frame as input and localizes the text regions (Fig. 8). Once the text is localized, the detected regions are fed to the script identification module to recognize the script of the detected text.

4.2 Script identification

Script identification takes text lines as input and identifies the script of the text. It is important to mention that during text detection, we treat it as a two-class problem, i.e., discrimination of text from non-text regions (irrespective of the script). It is also possible to treat it as a $k+1$ class problem where k represents the number of scripts ($k=2$ for our problem). In other words, the detection system, can be trained to detect the text in a particular script, hence avoiding the need of a separate script identification module. However, it is known that text in multiple scripts share some common characteristics (for instances edges of strokes, alignment, edge density etc.). Hence, designing a system to learn what is text and what is non-text seemed a more natural approach and is followed in our study.

For script identification, we employ CNNs in a classification framework (rather than detection). Urdu and English text lines are employed to fine-tune pre-trained ConvNets to discriminate between the two classes. Once trained, the model is able to separate text lines as a function of the script (Fig. 9).

Once the script of a text line is identified, it is fed to the respective recognition engine as discussed in the following.

4.3 Text recognition

As discussed earlier, we primarily target videos of News channels which contain cursive text (Urdu in our problem) along with text in the Roman script (English in our case). OCR systems for text in English (and other languages sharing the same script) are pretty mature. Hence, for recognition of English text, we employ off-the-shelf Google Tesseract OCR engine. For cursive text in Urdu, however, the performance of Google recognition engine was not very promising. Hence, we developed our own recognition engine (UrduNet) to recognize the text lines in Urdu. Recognition details are presented in the following.

4.3.1 Google Tesseract

Google Tesseract [10] is considered as the state-of-the-art OCR engine which provides high accuracy for many different languages including English. In our system, we have employed Tesseract version 4.0 which was recently released by Google. Version 4.0 is developed using deep neural network, and more specifically, it employs recurrent neural networks with long short-term memory architecture. The English text lines are fed to the recognition engine which returns the corresponding textual strings.

4.3.2 UrduNet

For recognition of Urdu text, we have designed and trained our own architecture which is a combination of a convolutional and a recurrent neural network and is termed as "UrduNet." The key motivation of employing a CNN is to convert raw pixel values into robust feature representations while a recurrent net is employed to model the problem using an implicit segmentation based approach. This allows directly feeding the text lines

Table 2 Experimental scenarios for text detection

		Traning		Testing	
Scenario	Dataset	Frames	Lines	Frames	Lines
S-I	Own Dataset	9546	53,274	4000	21,287
S-II	AcTiV [137]	1841	5143	4000	21,287
S-III	S-I + S-II	11,387	58,417	4000	21,287

Fig. 12 Text detection results on sample video frames

along with ground truth transcription to the model and no ligature or character level segmentation or labeling is required.

Recurrent nets have reported significant performance enhancements on problems like speech, handwriting, and caption text recognition, in the recent years. While simple RNNs fail to model long-term dependencies in the input sequences, variants like long short-term memory (LSTM) networks are commonly employed. LSTM represents a special type of recurrent unit with three gates, i.e., input, output, and forget. These gates are implemented using the sigmoid function and regulate the memory of an LSTM cell. It is also common to employ bi-directional LSTMs which parse the input in both forward and backward directions and concatenate the information for better predictions.

For feature extraction, we have designed a seven layer convolutional neural network. Input text line images are pre-processed and fed to the ConvNet. The pre-processing includes height-normalization, image binarization, and flipping. The flipping is carried out as Urdu is printed from right to left unlike western languages which are printed from left to right. Flipping the image ensures that the character sequences in the transcription are in correspondence with the image. The CNN maps input text line images to a feature map which is fed as a sequence to the recurrent layers. The recurrent part of the network is implemented using two layers of bidirectional LSTMs. The LSTM outputs pre-frame predictions which are converted to class labels using a CTC layer. Finally, a look-up table is used to map the class labels to the true Unicodes and produce the output transcription. A summary of these steps is illustrated in Fig. 10 while sample text lines used to train the model are presented in Fig. 11. Training is carried out in an end-to-end manner using the CTC loss function. Comprehensive findings on impact of pre-processing and the recurrent units can be found on our related studies [135] and [136], respectively.

Table 3 Text detection results on three experimental scenarios using different models

Models	Scenario	Precision	Recall	F-Measure
SSD	S-I	79.7	66.9	72.7
	S-II	75.3	56.1	64.3
	S-III	80.3	70.5	75.1
Faster R-CNN	S-I	83.1	84.9	84.0
	S-II	74.7	67.6	71.0
	S-III	**86.9**	**89.8**	**88.3**
East	S-I	78.2	45.3	57.3
	S-II	70.4	47.7	56.9
	S-III	77.3	47.9	59.1

5 Experiments, results, and discussion

To study the effectiveness of the proposed framework, a comprehensive series of experiments is carried out to evaluate the text detection, script identification, and text recognition modules. We first present the experimental protocol and realized results of text detection followed by

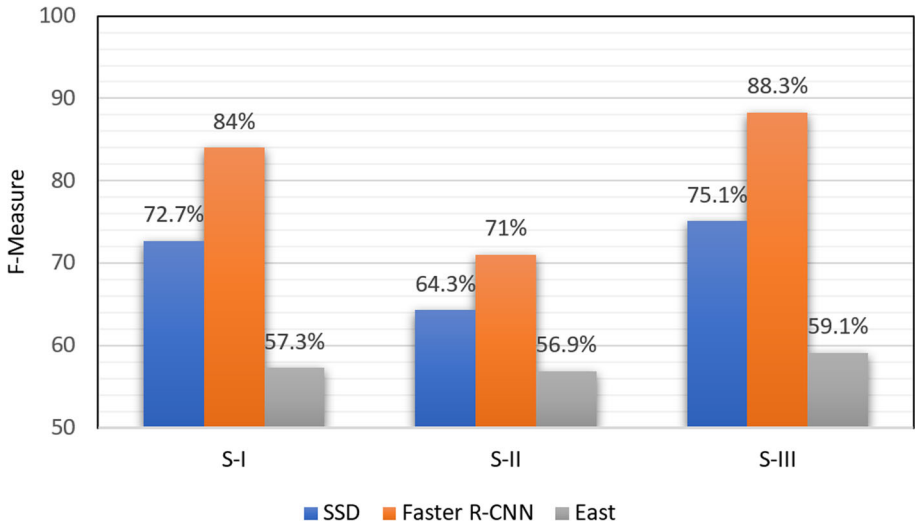

Fig. 13 Comparison of different models on detection performance

those of script identification. Finally, we discuss the performance of our recognition engine trained using different sets of images. In addition to our own dataset (presented in Section 3), we also employed other datasets in the training process as discussed in the following.

5.1 Text detection results

Text detection is evaluated by applying fine-tuning on different object detectors. These include Faster R-CNN [40], SSD [42] and EAST (Efficient and Accurate Scene Text detection pipeline) [49]. We used 9546 frames from our dataset containing more than 50,000 text lines for tuning these models while evaluations are carried out using more than 21,000 text lines from 4000 frames. Since Urdu and Arabic share many common characteristics, we also employed the publicly available dataset AcTiV developed by Zayene et al. [137]. The dataset contains about 1841 video frames with more than 5000 text lines, from different Arabic News channels. The text lines in the AcTiV dataset are used only in the training set and not in the test set. Three experimental scenarios are considered in our evaluation as listed in the following.

- **Scenario-I (S-I)**: Text lines from our custom-developed dataset are used to train the models.
- **Scenario-II (S-II)**: Text lines from AcTiV dataset [137] are used to train the models.
- **Scenario-III (S-III)**: The text lines in Scenario-I and Scenario-II are combined to train the models.

Table 2 summarizes the three scenarios along with the distribution of text lines into training and test sets for these scenarios.

The three models are trained for each of the scenarios by applying different settings of hyper parameters (optimized on the training set). Performance is quantified using the standard precision, recall, and F-measure where the bounding boxes of the detector are compared with

Fig. 14 Imperfect localization of text regions

Table 4 Text detection performance comparison with other techniques

Study	Language	Total frames	Precision	Recall	F-Measure
Jamail et al. [140]	Urdu	150	0.77	0.81	0.79
Siddiqi & Raza [141]	Urdu	1000	0.71	0.80	0.75
Moradi et al. [138]	Farsi/Arabic	4971	0.91	0.87	0.89
Raza et al. [142]	Urdu	1000	0.80	0.89	0.84
Raza et al. [142]	Arabic	300	0.81	0.93	0.86
Yousfi & Garcia [143]	Arabic	201	0.75	0.80	0.77
Zayene et al. [115]	Arabic	425	0.67	0.73	0.70
Zayene et al. [139]	Arabic	1843	0.83	0.85	0.84
Shahzad & Khurshid [144]	Urdu/Arabic	240	0.83	0.93	0.88
Mirza et al. [8]	Urdu	1000	0.72	0.89	0.80
Unar et al. [145]	Urdu	1000	0.83	0.88	0.85
Proposed method	**Urdu**	**13,645**	**0.87**	**0.89**	**0.88**

those in the ground truth. Detection results on sample video frames are illustrated in Fig. 12 while the quantified results are presented in Table 3. Comparing the performance of three models, it can be seen that Faster R-CNN outperforms the other two models. Comparing the three experimental scenarios, the lowest F-measures are reported in S-II which can be explained due to smaller number of (Arabic) text lines (around 5000) in the training set. Furthermore, the training set did not contain any text lines in English. The detection performance of S-I and S-III is comparable with S-III slightly better than S-I due to larger number of text lines in the training set. Over all, the highest F-measure of 88.3% is reported when using a combination of the two datasets in the training set. A comparison of detection performance of the three models for the three experimental scenarios is presented in Fig. 13.

In an attempt to provide an insight into the detection errors, few of the errors are illustrated in Fig. 14. It can be seen that in most cases, the detector is able to detect the textual region but the localization is not perfect, i.e., in some cases, the bounding box is larger (shorter) than the actual content leading to a reduced precision (recall).

Table 5 Distribution of data for script identification

	Traning		Testing	
	Frames	Lines	Frames	Lines
Urdu	9546	29,183	4000	11,287
English	9546	24,091	4000	10,000

Table 6 Confusion matrix for script identification

		Predicted	
		Urdu	English
Actual	Urdu	10,948	339
	English	395	9605

In order to compare the performance of our text detection method with already published works, we summarize the results of various studies targeting detection of cursive caption text in Table 4. It is important to mention that since different methods are evaluated on different datasets, a direct quantitative comparison may not be very meaningful. Most of the listed studies are evaluated on relatively smaller datasets (mostly $\leq$ 1000). Moradi et al. [138] and Zayene et al. [139] report results on relatively larger datasets with F-measures of 0.89 and 0.84, respectively. In comparison to other studies, we employ a significantly larger set of images with an F-measure of 0.88 indicating the robustness and scalability of our detector.

5.2 Script identification results

Script identification refers to classification of text regions into one of the script classes. In our study, we aim to distinguish between cursive Urdu and English text. Script identification is carried out by fine-tuning a pre-trained ConvNet with more than 29,000 Urdu and around 24,000 English text lines. For consistency, the distribution of data into training and test sets is kept the same as in case of experiments on text detection and is summarized in Table 5. The confusion matrix for script identification is presented in Table 6 where it can be seen that the model was able to recognize the scripts with less then 3% error rate.

5.3 Text recognition results

To study the recognition performance of our model (UrduNet), we carried out a series of experiments. It is important to mention that from the view point of a complete system, English text is recognized using off-the-shelf recognition engine. Consequently, we only report the

Table 7 Results of text recognition experiments

Experiment	Dataset	Training lines	Testing lines	Recognition rate (%)
1	Own dataset	29,183	11,287	83
2	UPTI [131]	10,000	11,287	45
3	Own dataset + UPTI	39,183	11,287	**87**

Fig. 15 Typical recognition errors made by UrduNet

recognition performance for Urdu text, the main theme of our study. In the first experiment, the model is trained using the text lines from our own dataset. We also train the model using 10,000 text lines of the UPTI [131] dataset of printed Urdu text to analyze how the system behaves once trained on printed text and evaluated on video text. Finally, we combine the text lines from video frames with those in the UPTI database to train the model using a huge set of more than 39,000 text lines. In all cases, the test set is kept same with more than 11,000 text lines. The (character) recognition rates realized in different experiments are summarized in Table 7. It can be seen that the once the model is trained using text lines from video frames, a recognition rate of 83% is reported. With UPTI text lines, the recognition rate drops significantly (46%). UPTI dataset contains high-resolution printed text lines, and once the system is only trained using these lines, the performance drops once tested on the challenging set of text lines from video frames. Combining the UPTI text lines with those from video frames (in the third experiment) reports the highest recognition rate of 87%. Few of the recognition errors are illustrated in Fig. 15 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 are correctly recognized in most cases.

We also provide a performance comparison of recent studies focusing on recognition of cursive text in general and Urdu text in particular. Since different methods are evaluated on different datasets under different experimental settings, a direct comparison of recognition rates is not the objective. The key idea is to provide an overview of the current state-of-the-art on this problem and assess the effectiveness of our recognizer with respect to it. Many interesting observations can be made from the summary of results presented in Table 8. The recognition rates on printed document images, in general, are naturally higher as compared to those reported on caption text. The highest reported recognition rate is 98.12% on 10,000 text lines of the UPTI dataset. It is however important to recall that UPTI contains synthetically generated text lines and do not offer the same kind of recognition challenges as those encountered in case of scanned documents or video text. In case of video text, Zayene et al. [114] reported 96.85% recognition rate of on a relatively smaller set of around 8000 Arabic text lines. For Urdu caption text, Tayyab et al. [134] reports 93% recognition rate on approximately 20,000 text lines. Hayat et al. [133] report a high classification rate of more than 99%, but the number of ligature classes is fairly small. In our experiments, we report a recognition rate of 87% which, though not directly comparable with reported studies, is indeed very promising

Table 8 Text detection performance comparison with other techniques

Image type	Study	Language	Data size	Recognition rate (%)
Document	Ahmed et al. [146]	Urdu	56 character clusters	93.40
	Akram et al. [147]	Urdu	224 images	86.15
	Hussain et al. [148]	Urdu	5249 ligatures	87.44
	Ahmed et al. [149]	Urdu	15,251 lines	96.00
	Naz et al. [123]	Urdu	10,000 lines	98.12
	Hassan et al. [150]	Urdu	10,000 lines	94.85
Videos	Zayene et al. [114]	Arabic	7843 lines	96.85
	Tayyab et al. [134]	Urdu	19,824 lines	93.02
	Hayat et al. [133]	Urdu	290 uniqe ligatures	99.50
	Proposed	**Urdu**	**40,470 lines**	**87.00**

considering the complexity of the problem and a much larger dataset.

6 Conclusion

In this paper, we presented a comprehensive framework for text detection and recognition in video frames containing textual occurrences in English and Urdu. A number of contributions are made in the presented study. We developed a comprehensive dataset of video frames with ground truth information allowing evaluation of detection and recognition tasks. For detection of textual regions, we employed state-of-the-art deep learning-based object detectors and fine-tuned them to detect text in multiple scripts. Script of the detected textual regions is identified using CNNs in a classification framework. A key contribution of this study is the development UrduNet, a combination of CNN and bidirectional LSTMs which reports high recognition rates for the challenging video text in cursive Urdu.

In our further work on this problem, we intend to develop a complete indexing and retrieval system that can be queried for keywords. The system will be optimized to work on live streams in addition to archived videos. This will in turn allow development of keyword based alert generation systems. Furthermore, the dataset is also planned to be enhanced and made available publicly. The study can also be extended to include additional scripts by integrating their respective OCRs.

Abbreviations

ANN: Artificial neural network; CBVR: Content-based video retrieval; CLE: Center of Language Engineering; CNN: Convolutional neural network; DCT: Discrete cosine transformation; DNN: Deep neural network; EAST: Efficient and Accurate Scene Text detection pipeline; GLCM: Gray level co-occurrence matrix; HMM: Hidden Markov models; HOG: Histogram of oriented gradients; LBP: Local binary pattern; LSTM: Long short-term memory; OCR: Optical character recognition; RNN: Recurrent neural network; R-CNN: Region-based convolutional neural network; SIFT: Scale invariant feature transform; SSD: Single Shot Detector; SVM: Support Vector Machine; UPTI: Urdu Printed Text Images; YOLO: You Only Look Once

Acknowledgements

Authors would like to thank IGNITE for funding this project.

Authors' contributions

Ali Mirza and Imran Siddiqi contributed to the algorithmic development and paper writing while Ossama Zeeshan and Muhammad Atif worked on implementation and experimental studies. All authors read and approved the final manuscript.

Funding

This research work is supported by IGNITE – National Technology Fund, Ministry of Information Technology, Government of Pakistan, under project number ICTRDF/TR&D/2014/35.

Availability of data and materials

Data will be made publicly available once the labeling process has completed.

Competing interests

The authors declare that they have no competing interests.

Received: 22 February 2019 Accepted: 13 August 2020
Published online: 28 August 2020

References

1. J. Burgess, J. Green, *YouTube: online video and participatory culture*. (Wiley, Cambridge, 2018)
2. R. Baran, P. Partila, R. Wilk, in *International Conference on Intelligent Human Systems Integration*. Automated text detection and character recognition in natural scenes based on local image features and contour processing techniques (Springer, Dubai, 2018), pp. 42–48
3. J. Dai, Z. Wang, X. Zhao, S. Shao, in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. Scene text detection based on enhanced multi-channels mser and a fast text grouping process (IEEE, Chengdu, 2018), pp. 351–355
4. S. Banerjee, K. Mullick, U. Bhattacharya, in *International Workshop on Camera-Based Document Analysis and Recognition*. A robust approach to extraction of texts from camera captured images (Springer, Washington, 2013), pp. 30–46
5. Y. Xu, S. Shan, Z. Qiu, Z. Jia, Z. Shen, Y. Wang, M. Shi, I. Eric, C. Chang, End-to-end subtitle detection and recognition for videos in east asian languages via cnn ensemble. Signal Process. Image Commun. **60**, 131–143 (2018)
6. T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, C. Sun, Single shot textspotter with explicit alignment and attention. arXiv preprint arXiv:1803.03474 (2018)
7. M. Liao, B. Shi, X. Bai, X. Wang, W. Liu. Textboxes: a fast text detector with a single deep neural network (AAAI, California, 2017), pp. 4161–4167
8. A. Mirza, M. Fayyaz, Z. Seher, I. Siddiqi, in *Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence*. Urdu caption text detection using textural features (ACM, Rabat, 2018), pp. 70–75
9. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, in *CVPR, vol. 2*. Learning from simulated and unsupervised images through adversarial training (IEEE, Honolulu, 2017), p. 5
10. R. Smith, in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference On, vol. 2*. An overview of the Tesseract OCR engine (IEEE, Paraná, 2007), pp. 629–633
11. Y. Qiaoyang, D. Doermann., Text detection and recognition in imagery: a survey. IEEE Trans. Patt. Anal. Mach. Intell. **37**, 1480–1500 (2015)
12. S. Wang, C. Fu, Q. Li, in *International Conference on Machine Learning and Intelligent Communications*. Text detection in natural scene image: a survey (Springer, Shanghai, 2016), pp. 257–264
13. X.-C. Yin, Z.-Y. Zuo, S. Tian, C.-L. Liu, Text detection, tracking and recognition in video: a comprehensive survey. IEEE Trans. Image Process. **25**(6), 2752–2773 (2016)
14. H. Zhang, K. Zhao, Y.-Z. Song, J. Guo, Text extraction from natural scene image: a survey. Neurocomputing. **122**, 310–323 (2013)
15. N. Sharma, U. Pal, M. Blumenstein, in *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop On*. Recent advances in video based document processing: a review (IEEE, Gold Coast, 2012), pp. 63–68
16. A. Jamil, I. Siddiqi, F. Arif, A. Raza, in *International Conference on Document Analysis and Recognition*. Edge-based features for localization of artificial Urdu text in video images (IEEE, Beijing, 2011)
17. Y. C. Kiran, L. N. C, Text extraction and verification from video based on SVM. World J. Sci. Technol. **2**(5), 124–126 (2012)
18. Y.-F. Pan, X. Hou, C.-L. Liu, A hybrid approach to detect and localize texts in natural scene images. IEEE Trans. Image Process. **20**(3), 800–813 (2011)
19. X. Bai, C. Yao, W. Liu, Strokelets: a learned multi-scale mid-level representation for scene text recognition. IEEE Trans. Image Process. **25**, 2789–2802 (2016)
20. X. Huang, in *4th International Congress on Image and Signal Processing*. A novel video text extraction approach based on log-Gabor filters (IEEE, Shanghai, 2011)
21. D. Gabor, J. Inst. Electr. Eng.- III Radio Commun. Eng. **93**(26), 429–441 (1946)
22. Q. Ye, Q. Huang, W. Gao, D. Zhao, Fast and robust text detection in images and video frames. Image Vis. Comput. **23**(6), 565–576 (2005)
23. J. Guillaume, E. Véronique, B. Stéphane, E. Hubert, in *Electronic Imaging 2007*. Curvelets based feature extraction of handwritten shapes for ancient manuscripts classification (International Society for Optics and Photonics, California, 2007), pp. 65000–65000
24. M. Anthimopoulos, B. Gatos, I. Pratikakis, A two-stage scheme for text detection in video images. Image and Vis. Comput. **28**(9), 1413–1426 (2010)
25. Y. Zhong, H. Zhang, A. K. Jain, Automatic caption localization in compressed video. IEEE Trans. Patt. Anal. Mach. Intell. **22**(4), 385–392 (2000)

26. N. Dalal, B. Triggs, in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference On, vol. 1*. Histograms of oriented gradients for human detection (IEEE, San Diego, 2005), pp. 886–893
27. P. Shivakumara, T. Q. Phan, C. L. Tan, New fourier-statistical features in RGB space for video text detection. IEEE Trans. Circ. Syst. Video Technol. **20**(11), 1520–1532 (2010)
28. C. Yi, Y. Tian, Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification. IEEE Trans. Image Process. **21**(9), 4256–4268 (2012)
29. P. Shivakumara, T. Q. Phan, C. L. Tan, New fourier-statistical features in RGB space for video text detection. IEEE Trans. Circuits Syst. Video Technol. **20**, 1520–1532 (2010)
30. P. Shivakumara, R. P. Sreedhar, T. Q. Phan, S. Lu, C. L. Tan, Multioriented video scene text detection through Bayesian classification and boundary growing. IEEE Trans. Circ. Syst. Video Technol. **22**(8), 1227–1235 (2012)
31. W. Zhen, W. Zagiqiang, in *2nd International Symposium on Computational Intelligence and Design*. A comparative study of feature selection for SVM in video text detection (IEEE, Changsha, 2009), pp. 552–556
32. X.-C. Yin, X. Yin, K. Huang, Robust text detection in natural scene images. IEEE Trans. Patt. Anal. Mach. Intell. **36**(5), 970–983 (2013)
33. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature. **521**(7553), 436 (2015)
34. A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Advances in Neural Information Processing Systems*. Imagenet classification with deep convolutional neural networks (Neural Information Processing Systems Foundation, Inc. (NIPS), Nevada, 2012), pp. 1097–1105
35. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al*, Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
36. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al., *Going deeper with convolutions*. (Cvpr, Boston, 2015)
38. R. Girshick, J. Donahue, T. Darrell, J. Malik, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Rich feature hierarchies for accurate object detection and semantic segmentation (IEEE, Columbus, 2014), pp. 580–587
39. R. Girshick, in *Proceedings of the IEEE International Conference on Computer Vision*. Fast R-CNN (IEEE, Las Condes, 2015), pp. 1440–1448
40. S. Ren, K. He, R. Girshick, J. Sun, in *Advances in Neural Information Processing Systems*. Faster R-CNN: towards real-time object detection with region proposal networks (Neural Information Processing Systems Foundation, Inc. (NIPS), Quebec, 2015), pp. 91–99
41. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. You only look once: unified, real-time object detection (IEEE, Las Vegas, 2016), pp. 779–788
42. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, in *European Conference on Computer Vision*. SSD: single shot multibox detector (Springer, Amsterdam, 2016), pp. 21–37
43. W. Huang, Y. Qiao, X. Tang, in *European Conference on Computer Vision*. Robust scene text detection with convolution neural network induced MSER trees (Springer, Zürich, 2014), pp. 497–511
44. Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, X. Bai, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Multi-oriented text detection with fully convolutional networks (IEEE, Las Vegas, 2016), pp. 4159–4167
45. A. Gupta, A. Vedaldi, A. Zisserman, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Synthetic data for text localisation in natural images (IEEE, Las Vegas, 2016), pp. 2315–2324
46. B. Shi, X. Bai, S. Belongie, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Detecting oriented text in natural images by linking segments, (2017), pp. 2550–2558
47. Z. Tian, W. Huang, T. He, P. He, Y. Qiao, in *European Conference on Computer Vision*. Detecting text in natural image with connectionist text proposal network (Springer, Amsterdam, 2016), pp. 56–72
48. Y. Wang, C. Shi, B. Xiao, C. Wang, C. Qi, CRF based text detection for natural scene images using convolutional neural network and context information. Neurocomputing. **295**, 46–58 (2018)
49. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, in *Proc. CVPR*. East: an efficient and accurate scene text detector (IEEE, Hawaii, 2017), pp. 2642–2651
50. K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo, T. Yibulayin, Script identification of multi-script documents: a survey. IEEE Access. **5**, 6546–6559 (2017)
51. D. Ghosh, T. Dube, A. Shivaprasad, Script recognition—a review. IEEE Trans. Patt. Anal. Mach. Intell. **32**(12), 2142–2161 (2010)
52. M. Li, M. Bai, in *Intelligent Control and Automation (WCICA), 2012 10th World Congress On*. A mixed edge based text detection method by applying image complexity analysis (IEEE, Beijing, 2012), pp. 4809–4814
53. A. Jamil, A. Batool, Z. Malik, A. Mirza, I. Siddiqi, Multilingual artificial text extraction and script identification from video images. Int. J. Adv. Comput. Sci. Appl. **1**(7), 529–539 (2016)
54. J. Hochberg, L. Kerns, P. Kelly, T. Thomas, in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference On, vol. 1*. Automatic script identification from images using cluster-based templates (IEEE, Montreal, 1995), pp. 378–381
55. A. L. Spitz, Determination of the script and language content of document images. IEEE Trans. Patt. Anal. Mach. Intell. **19**(3), 235–245 (1997)
56. U. Pal, B. Chaudhuri, in *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference On*. Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line (IEEE, Washington, 2001), pp. 790–794
57. Z. Malik, A. Mirza, A. Bennour, I. Siddiqi, C. Djeddi, in *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. Video script identification using a combination of textural features (IEEE, Bangkok, 2015), pp. 61–67
58. C. Zhu, W. Wang, Q. Ning, in *Advances in Multimedia Information Processing-PCM 2006*. Text detection in images using texture feature from strokes (Springer, Hangzhou, 2006), pp. 295–301
59. Z. Li, G. Liu, X. Qian, D. Guo, H. Jiang, Effective and efficient video text extraction using key text points. IET Image Process. **5**(8), 671–683 (2011)
60. N. Sharma, S. Chanda, U. Pal, M. Blumenstein, in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference On*. Word-wise script identification from video frames (IEEE, Washington, 2013), pp. 867–871
61. G. Peake, T. Tan, in *Document Image Analysis, 1997.(DIA'97) Proceedings., Workshop On*. Script and language identification from document images (IEEE, San Juan, 1997), pp. 10–17
62. D. Zhao, P. Shivakumara, S. Lu, C. L. Tan, New spatial-gradient-features for video script identification. Doc. Anal. Syst., 38–42 (2012)
63. P. Shivakumara, N. Sharma, U. Pal, M. Blumenstein, C. L. Tan, in *Pattern Recognition (ICPR), 2014 22nd International Conference On*. Gradient-angular-features for word-wise video script identification (IEEE, Stockholm, 2014), pp. 3098–3103
64. N. Sharma, U. Pal, M. Blumenstein, in *Neural Networks (IJCNN), 2014 International Joint Conference On*. A study on word-level multi-script identification from video frames (IEEE, Beijing, 2014), pp. 1827–1833
65. N. Sharma, R. Mandal, R. Sharma, U. Pal, M. Blumenstein, in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference On*. Icdar2015 competition on video script identification (CVSI 2015) (IEEE, Nancy, 2015), pp. 1196–1200
66. J. Mei, L. Dai, B. Shi, X. Bai, in *2016 23rd International Conference on Pattern Recognition (ICPR)*. Scene text script identification with convolutional recurrent neural networks (IEEE, Cancún, 2016), pp. 4053–4058
67. A. K. Singh, A. Mishra, P. Dabral, C. Jawahar, in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop On*. A simple and effective solution for script identification in the wild (IEEE, Santorini, 2016), pp. 428–433
68. L. Gomez, D. Karatzas, in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop On*. A fine-grained approach to scene text script identification (IEEE, Santorini, 2016), pp. 192–197
69. L. Gomez, A. Nicolaou, D. Karatzas, Improving patch-based scene text script identification with ensembles of conjoined networks. Patt. Recogn. **67**, 85–96 (2017)
70. M. Tounsi, I. Moalla, F. Lebourgeois, A. M. Alimi, in *International Conference on Neural Information Processing*. CNN based transfer learning for scene script identification (Springer, California, 2017), pp. 702–711
71. J. Zdenek, H. Nakayama, in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference On, vol. 1*. Bag of local convolutional triplets for script identification in scene text (IEEE, Kyoto, 2017), pp. 369–375
72. A. K. Bhunia, A. Konwer, A. K. Bhunia, A. Bhowmick, P. P. Roy, U. Pal, Script identification in natural scene image and video frames using an attention based convolutional-LSTM network. Patt. Recogn. **85**, 172–184 (2019)

73. S. Palaiahnakote, P. T. Quy, T. C. Lim, A Laplacian approach to multi-oriented text detection in video. IEEE Trans. Patt. Anal. Mach. Intell. **33**(2), 412–419 (2011)
74. A. Garz, M. Seuret, F. Simistira, A. Fischer, R. Ingold, in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop On*. Creating ground truth for historical manuscripts with document graphs and scribbling interaction (IEEE, Santorini, 2016), pp. 126–131
75. J.-B. Fasquel, N. Delanoue, A graph based image interpretation method using a priori qualitative inclusion and photometric relationships. IEEE Trans. Patt. Anal. Mach. Intell. **41**, 1043–1055 (2018)
76. N. Islam, Z. Islam, N. Noor, A survey on optical character recognition system. arXiv preprint arXiv:1710.05703 (2017)
77. B. Lei, G. Xu, M. Feng, F. Van der Heijden, Y. Zou, D. de Ridder, D. M. Tax, *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*. (Wiley, 2017)
78. A. H. Metwally, M. I. Khalil, H. M. Abbas, in *Computer Engineering and Systems (ICCES), 2017 12th International Conference On*. Offline arabic handwriting recognition using hidden Markov models and post-recognition lexicon matching (IEEE, Cairo, 2017), pp. 238–243
79. M. Rabi, M. Amrouch, Z. Mahani, Int. J. Pattern Recognit. Artif. Intell. **32**(01), 1860007 (2018)
80. M. Rabi, M. Amrouch, Z. Mahani, Cursive arabic handwriting recognition system without explicit segmentation based on hidden Markov models. J. Data Min. Digit. Human. (2018)
81. . Caner, I. Haritaoglu, in *Pattern Recognition (ICPR), 2010 20th International Conference On*. Shape-DNA: effective character restoration and enhancement for Arabic text documents (IEEE, Istanbul, 2010), pp. 2053–2056
82. P. S. Kompalli, *Image document processing in a client-server system including privacy-preserving text recognition*. (Google Patents, 2017). US Patent 9,847,974
83. V. Märgner, U. Pal, A. Antonacopoulos, et al., Document analysis and text recognition. World Sci. (2018)
84. S. Sudholt, G. A. Fink, in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference On*. PHOCNet: a deep convolutional neural network for word spotting in handwritten documents (IEEE, 2016), pp. 277–282
85. C.-L. Liu, G. A. Fink, V. Govindaraju, L. Jin, *Special issue on deep learning for document analysis and recognition*. (Springer, 2018)
86. T. Quy Phan, P. Shivakumara, S. Tian, C. Lim Tan, in *Proceedings of the IEEE International Conference on Computer Vision*. Recognizing text with perspective distortion in natural scenes (IEEE, Sydney, 2013), pp. 569–576
87. C. Yi, Y. Tian, Scene text recognition in mobile applications by character descriptor and structure configuration. IEEE Trans. Image Process. **23**(7), 2972–2982 (2014)
88. C. Yao, X. Bai, B. Shi, W. Liu, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Strokelets: a learned multi-scale representation for scene text recognition (IEEE, Columbus, 2014), pp. 4042–4049
89. S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, C. L. Tan, Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. Patt. Recogn. **51**, 125–134 (2016)
90. B. Yu, H. Wan, Chinese text detection and recognition in natural scene using hog and SVM. DEStech Trans. Comput. Sci. Eng. (2016)
91. M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
92. S. Kumar, K. Kumar, R. K. Mishra, et al., Scene text recognition using artificial neural network: a survey. Int. J. Comput. Appl. **137**(6), 40–50 (2016)
93. S. Zhu, R. Zanibbi, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. A text detection system for natural scenes with convolutional feature learning and cascaded classification (IEEE, Nevada, 2016), pp. 625–632
94. S. Lu, T. Chen, S. Tian, J.-H. Lim, C.-L. Tan, Scene text extraction based on edges and support vector regression. Int. J. Doc. Anal. Recogn. (IJDAR). **18**(2), 125–135 (2015)
95. L. Neumann, Scene text localization and recognition in images and videos (2017). Department of cybernetics Faculty of Electrical Engineering, Czech Technical University
96. V. Goel, A. Mishra, K. Alahari, C. Jawahar, in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference On*. Whole is greater than sum of parts: recognizing scene text words (IEEE, Washington, 2013), pp. 398–402
97. M. Jaderberg, A. Vedaldi, A. Zisserman, in *European Conference on Computer Vision*. Deep features for text spotting (Springer, Zürich, 2014), pp. 512–528
98. A. Salhi, B. Minaoui, M. Fakir, H. Chakib, H. Grimech, Traffic signs recognition using HP and HOG descriptors combined to MLP and SVM classifiers. Traffic. **8**(11) (2017)
99. K. C. Saranya, V. Singhal, in *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*. Real-time prototype of driver assistance system for indian road signs (Springer, 2018), pp. 147–155
100. Y. Lai, N. Wang, Y. Yang, L. Lin, in *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM), Madeira, Portugal*. Traffic signs recognition and classification based on deep feature learning (Springer, Madeira, 2018), pp. 622–629
101. B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Robust scene text recognition with automatic rectification (IEEE, Nevada, 2016), pp. 4168–4176
102. B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Patt. Anal. Mach. Intell. **39**(11), 2298–2304 (2017)
103. H. Yang, S. Li, X. Yin, A. Han, J. Zhang, in *Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference On*. Recurrent highway networks with attention mechanism for scene text recognition (IEEE, Sydney, 2017), pp. 1–8
104. M. Bušta, L. Neumann, J. Matas, in *Computer Vision (ICCV), 2017 IEEE International Conference On*. Deep textspotter: an end-to-end trainable scene text localization and recognition framework (IEEE, Venice, 2017), pp. 2223–2231
105. Z. Lei, S. Zhao, H. Song, J. Shen, Scene text recognition using residual convolutional recurrent neural network. Mach. Vis. Appl. **29**, 1–11 (2018)
106. Z. Liu, Y. Li, F. Ren, W. L. Goh, H. Yu, in *AAAI*. SqueezedText: a real-time scene text recognition by binary convolutional encoder-decoder network (AAAI, Louisiana, 2018)
107. W. Liu, C. Chen, K.-Y. K. Wong, in *AAAI*. Char-Net: a character-aware neural network for distorted scene text recognition (AAAI, Louisiana, 2018)
108. M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, X. Bai, in *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33*. Scene text recognition from two-dimensional perspective, (2019), pp. 8714–8721
109. Y. Gao, Z. Huang, Y. Dai, Double supervised network with attention mechanism for scene text recognition. arXiv preprint arXiv:1808.00677 (2018)
110. D. Kim, K. Sohn, in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference On*. Static text region detection in video sequences using color and orientation consistencies (IEEE, Florida, 2008), pp. 1–4
111. T. Q. Phan, P. Shivakumara, T. Lu, C. L. Tan, in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference On*. Recognition of video text through temporal integration (IEEE, Washington, 2013), pp. 589–593
112. P. Kulkarni, P. Bhagyashri, B. Joglekar, in *Industrial Instrumentation and Control (ICIC), 2015 International Conference on. IEEE*. An effective content based video analysis and retrieval using pattern indexing techniques (IEEE, Pune, 2015)
113. T. A. N., C. Vaidya, P. Dahiwale, in *Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on. IEEE*. A survey of indexing techniques for large scale content-based image retrieval (IEEE, Visakhapatnam, 2015)
114. O. Zayene, S. M. Touj, J. Hennebert, R. Ingold, N. E. B. Amara, Multi-dimensional long short-term memory networks for artificial arabic text recognition in news video. IET Comput. Vis. **12**, 710–719 (2018)
115. O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, N. E. B. Amara, in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. A dataset for arabic text detection, tracking and recognition in news videos - activ, (2015)
116. S. Yousfi, S.-A. Berrani, C. Garcia, in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference On*. Alif: a dataset for arabic embedded text recognition in tv broadcast (IEEE, Nancy, 2015), pp. 1221–1225
117. S. Yousfi, S.-A. Berrani, C. Garcia, in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference On*. Deep learning and recurrent connectionist-based approaches for arabic text recognition in videos (IEEE, Nancy, 2015), pp. 1026–1030

118. M. Jain, M. Mathew, C. Jawahar, in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. Unconstrained scene text and video text recognition for arabic script, (IEEE, 2017), pp. 26–30
119. W. Lu, H. Sun, J. Chu, X. Huang, J. Yu, A novel approach for video text detection and recognition based on a corner response feature map and transferred deep convolutional neural network. IEEE Access. **6**, 40198–40211 (2018)
120. H. Malik, M. A. Fahiem, in *2009 Second International Conference in Visualisation*. Segmentation of printed urdu scripts using structural features (IEEE, 2009), pp. 191–195
121. S. Naz, A. I. Umar, R. Ahmad, S. B. Ahmed, S. H. Shirazi, I. Siddiqi, M. I. Razzak, Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks. Neurocomputing. **177**, 228–241 (2016)
122. N. Javed, S. Shabbir, I. Siddiqi, K. Khurshid, in *Frontiers of Information Technology (FIT), 2017 International Conference On*. Classification of Urdu ligatures using convolutional neural networks-a novel approach (IEEE, Islamabad, 2017), pp. 93–97
123. S. Naz, A. I. Umar, R. Ahmad, I. Siddiqi, S. B. Ahmed, M. I. Razzak, F. Shafait, Urdu Nastaliq recognition using convolutional–recursive deep learning. Neurocomputing. **243**, 80–87 (2017)
124. I. Ahmad, S. A. Mahmoud, G. A. Fink, Open-vocabulary recognition of machine-printed Arabic text using hidden Markov models. Patt. Recogn. **51**, 97–111 (2016)
125. A. Khemiri, A. K. Echi, A. Belaid, M. Elloumi, in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference On*. Arabic handwritten words off-line recognition based on HMMs and DBNs (IEEE, Nancy, 2015), pp. 51–55
126. S. T. Javed, S. Hussain, in *Iberoamerican Congress on Pattern Recognition*. Segmentation based Urdu Nastalique OCR (Springer, Havana, 2013), pp. 41–49
127. I. U. Din, I. Siddiqi, S. Khalid, T. Azam, Segmentation-free optical character recognition for printed Urdu text. EURASIP J. Image Video Process. **2017**(1), 62 (2017)
128. G. S. Lehal, in *Proceeding of the Workshop on Document Analysis and Recognition*. Choice of recognizable units for Urdu OCR (ACM, Mumbai, 2012), pp. 79–85
129. S. Hassan, A. Irfan, A. Mirza, I. Siddiqi, in *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*. Cursive handwritten text recognition using bi-directional LSTMs: a case study on urdu handwriting (IEEE, Istanbul, 2019), pp. 67–72
130. S. Naz, A. I. Umar, R. Ahmed, M. I. Razzak, S. F. Rashid, F. Shafait, Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks. SpringerPlus. **5**(1), 2010 (2016)
131. N. Sabbour, F. Shafait, in *IS&T/SPIE Electronic Imaging*. A segmentation-free approach to Arabic and Urdu OCR (International Society for Optics and Photonics, 2013), pp. 86580–86580
132. CLE, Center for Language Engineering. http://http://www.cle.org.pk/. Accessed 10 Oct 2018
133. U. Hayat, M. Aatif, O. Zeeshan, I. Siddiqi, in *2018 14th International Conference on Emerging Technologies (ICET)*. Ligature recognition in Urdu caption text using deep convolutional neural networks (IEEE, Islamabad, 2018), pp. 1–6
134. B. U. Tayyab, M. F. Naeem, A. Ul-Hasan, F. Shafait, et al., in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. A multi-faceted OCR framework for artificial Urdu news ticker text recognition (IEEE, Vienna, 2018), pp. 211–216
135. A. Mirza, I. Siddiqi, S. G. Mustufa, M. Hussain, in *Iberian Conference on Pattern Recognition and Image Analysis*. Impact of pre-processing on recognition of cursive video text (Springer, Madrid, 2019), pp. 565–576
136. A. Mirza, I. Siddiqi, Recognition of cursive video text using a deep learning framework. IET Image Process. (2020)
137. Z. Oussama, H. Jean, T. S. Masmoudi, I. Rolf, A. N. E. Ben, in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference On*. A dataset for arabic text detection, tracking and recognition in news videos-AcTiV (IEEE, Nancy, 2015), pp. 996–1000
138. M. Moradi, S. Mozaffari, Hybrid approach for Farsi/Arabic text detection and localisation in video frames. IET Image Process. **7**(2), 154–164 (2013)
139. O. Zayene, J. Hennebert, M. Seuret, S. M. Touj, R. Ingold, N. E. B. Amara, *Text detection in arabic news video based on SWT operator and convolutional auto-encoders* (IEEE, Santorini, 2016)
140. A. Jamil, I. Siddiqi, F. Arif, A. Raza, in *Document Analysis and Recognition (ICDAR), 2011 International Conference On*. Edge-based features for localization of artificial Urdu text in video images (IEEE, Peking, 2011), pp. 1120–1124
141. I. Siddiqi, A. Raza, in *MMEDIA 2012, The Fourth International Conferences on Advances in Multimedia*. A database of artificial urdu text in video images with semi-automatic text line labeling scheme (IARIA XPS Press, Mont Blanc, 2012), pp. 75–81
142. A. Raza, I. Siddiqi, C. Djeddi, A. Ennaji, in *2013 12th International Conference on Document Analysis and Recognition*. Multilingual artificial text detection using a cascade of transforms (IEEE, Washington, 2013), pp. 309–313
143. S. Yousfi, S.-A. Berrani, C. Garcia, in *2014 IEEE International Conference on Image Processing (ICIP)*. Arabic text detection in videos using neural and boosting-based approaches: application to video indexing (IEEE, Paris, 2014), pp. 3028–3032
144. U. Shahzad, K. Khurshid, in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. Oriental-script text detection and extraction in videos (IEEE, Nancy, 2017), pp. 15–20
145. S. Unar, A. H. Jalbani, M. M. Jawaid, M. Shaikh, A. A. Chandio, Artificial urdu text detection and localization from individual video frames. Mehran Univ. Res. J. Eng. Technol. **37**(2), 429–438 (2018)
146. Z. Ahmad, J. K. Orakzai, I. Shamsher, A. Adnan, in *Proc. of World Academy of Science, Engineering and Technology*. Urdu Nastaleeq optical character recognition (World Academy of Science, Engineering and Technology (WASET), Paris, 2007), pp. 249–252
147. Q. Akram, S. Hussain, F. Adeeba, S. Rehman, M. Saeed, in *Proc. of Conference on Language and Technology (CLT)*. Framework of Urdu Nastalique optical character recognition system, (Karachi, 2014), pp. 1–7
148. S. Hussain, S. Ali, Q. u. a. Akram, Nastalique segmentation-based approach for Urdu OCR. Int. J. Doc. Anal. Recogn. (IJDAR). **18**(4), 357–374 (2015)
149. S. B. Ahmed, S. Naz, M. I. Razzak, S. F. Rashid, M. Z. Afzal, T. M. Breuel, Evaluation of cursive and non-cursive scripts using recurrent neural networks. Neural Comput. Appl. **27**(3), 603–613 (2016)
150. A. Ul-Hasan, S. B. Ahmed, F. Rashid, F. Shafait, T. M. Breuel, in *2013 12th International Conference on Document Analysis and Recognition*. Offline printed Urdu Nastaleeq script recognition with bidirectional LSTM networks (IEEE, Washington, 2013), pp. 1061–1065

Publisher's Note