applied sciences

MDPI

Article

Arabic Cursive Text Recognition from Natural Scene Images

Saad Bin Ahmed [1,2,*], Saeeda Naz [3], Muhammad Imran Razzak [4] and Rubiyah Yusof [1]

[1] Malaysia-Japan International Institute of Technology (M-JIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia; rubiyah.kl@utm.my
[2] Department of Health Informatics, King Saud bin Abdulaziz University for Health Sciences, Ministry of National Guard Health Affairs (NGHA), Riyadh 11481, Saudi Arabia
[3] GPGC No. 1, Higher Education Department, Abbottabad 22010, Pakistan; saeedanaz292@gmail.com
[4] Department of Information Technology, University of Technology Sydney, Sydney 2007, Australia; imran.razzak@ieee.org
* Correspondence: isaadahmed@gmail.com or saad2@utm.my

Received: 29 November 2018; Accepted: 31 December 2018; Published: 10 January 2019

Abstract: This paper presents a comprehensive survey on Arabic cursive scene text recognition. The recent years' publications in this field have witnessed the interest shift of document image analysis researchers from recognition of optical characters to recognition of characters appearing in natural images. Scene text recognition is a challenging problem due to the text having variations in font styles, size, alignment, orientation, reflection, illumination change, blurriness and complex background. Among cursive scripts, Arabic scene text recognition is contemplated as a more challenging problem due to joined writing, same character variations, a large number of ligatures, the number of baselines, etc. Surveys on the Latin and Chinese script-based scene text recognition system can be found, but the Arabic like scene text recognition problem is yet to be addressed in detail. In this manuscript, a description is provided to highlight some of the latest techniques presented for text classification. The presented techniques following a deep learning architecture are equally suitable for the development of Arabic cursive scene text recognition systems. The issues pertaining to text localization and feature extraction are also presented. Moreover, this article emphasizes the importance of having benchmark cursive scene text dataset. Based on the discussion, future directions are outlined, some of which may provide insight about cursive scene text to researchers.

Keywords: scene text recognition; Arabic cursive scripts; supervised learning; natural scene images; text recognition

1. Introduction

Advancement in cameras of hand-held gadgets prompt their users to capture scene images having overlaid text. In today's era, most people have specialized gadgets to capture scene images for obtaining information during work, a journey, etc. The camera-captured images may contain much textual information in addition to semantic knowledge represented by graphics or in pictures. The text appearing in natural images is usually used for conveying information to people. The scene text is represented by different types of font styles and sizes having various backgrounds, including building, sea, mountain, forest, etc., which are termed as noise and may halt the smooth process of text recognition from natural images. The natural images having text can be seen on signboards, banners and advertising notes or boards. The text extraction from a natural image is an emerging research field as far as cursive scene text recognition is concerned. This problem is noted to be a challenging one due to implicit noise like blur, lighting condition, text alignment, styles, orientation and the size of text attached to the image. Moreover, a complex background sometimes makes it difficult to extract the text in comparison with traditional Optical Character Recognition (OCR).

The extracted text from natural scene images is beneficial for applications such as text search in a video, text extraction from videos, content-based retrieval, search engines and in those applications where text in an image has an important concern. The focus of this paper remains on the investigation and analysis of state-of-the-art techniques developed for text extraction and recognition of textual data as captured in an image. This paper particularly emphasizes the work presented in recently-proposed research on cursive text, especially Arabic scene text recognition.

A cursive word is more related to and used with script, which means any penmanship style by using the various symbols of any language that is written in a conjoined and flowing way. The text other than printed Latin may be taken from different writing styles. There are numerous complicated scripts that exist, like in Arabic, Chinese, and Japanese, that are categorized as cursive in nature, either taken by a specialized camera, synthetic means or in handwritten form. Plain or Latin text recognition is no longer a research problem as many researchers have proposed a solution for efficient printed, handwritten and scene text recognition for Latin [1–4] systems. However, the recognition of non-Latin scripts still poses a great challenge and requires more effort from the research community to address it.

A non-Latin and prominent cursive script is Arabic. In Arabic, the text is cursive in nature because isolated characters do not represent any meaning unless they are used in conjunction with other characters, as presented in Figure 1.

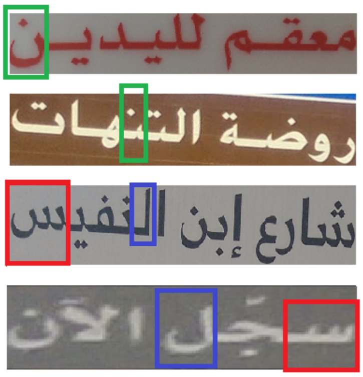


Figure 1. Arabic as cursive script. The green box shows different positions of the Arabic character noon, while the blue box represents different positions of the Arabic character laam. The red box depicts various positions of the Arabic character seen.

There are four positions of a character in a word. These positions may be initial, middle, final or as an isolated character. Due to various appearances of the same character, the segmentation becomes very difficult to perform. For this reason, implicit segmentation approaches are presented to handle the segmentation problem [4,5]. The following are the highlighted issues related to Arabic-like cursive scripts.

1. More than one shape of a character increases the complexity in character recognition.
2. The diacritics on a character are a more important feature of some characters in Arabic, because without using them, it becomes difficult to read a proper word and understand its meaning.
3. Contrary to Latin, the writing style is from right to left.

In recent years, the research work based on implicit segmentation and context learning classifiers on Arabic/Urdu script OCRs either in printed or in handwritten form have been described in [4–6]. The state-of-the-art technique like Recurrent Neural Networks (RNNs) [7] has been applied to Urdu cursive script and resulted in achieving remarkable accuracies as reported on this intrinsic script in [4,5,8]. Although the work on optical character recognition of Urdu-Arabic-like scripts tried to present commendable solutions as shown in recent research publication on this subject, the recognition of Arabic scene text has not shown significant results yet. As discussed earlier, the techniques that have been applied to OCR systems have failed catastrophically on Arabic scene text recognition. This is due to the complex structure of Arabic scene text images in the presence of varied styles of text represented in any colour and in any orientation regardless of following any font style and size. Efforts are being made to overcome the difficulties in this direction, and some research work has been reported, which will be detailed later in this article. There is supplementary inter- and intra-class variation in the text extracted from natural images. It is comparatively easier to recognize Latin script from natural images, unlike Arabic, Chinese, Japanese or any other cursive script. Figure 2 shows scene text images of different languages.

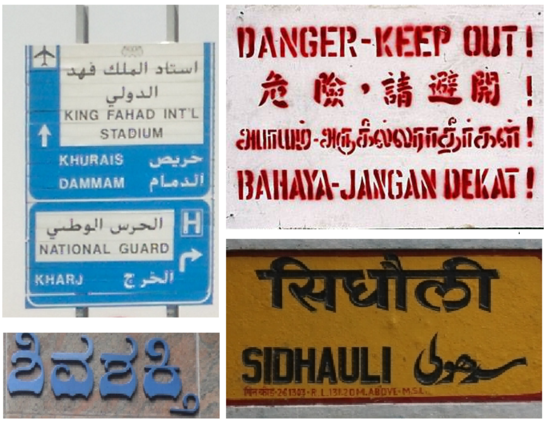


Figure 2. Multilingual scene text images. The text in the images is representative of Chinese, Hindi, Urdu, English and Tamil.

This survey summarizes the work of other researchers who have contributed to cursive scene text recognition since 2009 to date. In addition to that, the status of Arabic or Arabic-like scripts is also accentuated. Arabic script is one of the most common and the second largest language, having the status of the national languages of the Arabian Peninsula. Around more than one billion users in the world communicate in the Arabic script-based languages in reading and writing. The writing style of Arabic script is from right to left with the combination of diacritics, which is considered an integral part in making a word more meaningful. There is a variety of techniques presented for Arabic or Arabic-like text recognition, either in printed [8], scanned or handwritten format [6].

The existing methods designed for scene text detection and recognition may be categorized into texture-based, component-based and hybrid-based methods.

1. The texture-based method relies on the properties of an image like intensity and hue values, wavelet transformation of an image and by applying different filtration techniques, which contribute to representing the image. Such properties may help to detect the text in an image as explained in some of the presented work, like [9–12].
2. The component-based method depends on the specific region(s) of an image. The region is often marked by colour clustering and coordinate values. Different filtration techniques may be applied to segment the text and non-text region from an image. If scene text images are taken in specific settings, then component-based methods produce good results. This method is not suitable for invariant text images like the difference in font size, rotation, etc. Some researchers proposed their techniques by using this method, as mentioned by the researchers in [4,5,7,13,14].
3. The hybrid methods share the characteristics of both texture-based and component-based methods. The candidate regions is determined by using both techniques on the same image, as explored in some works, as mentioned in [15–18].

The text segmentation approaches that have been applied to OCR systems can also be applied to scene text recognition systems, as reported by [19,20]. These approaches produced state-of-the-art results on OCR either being applied to cursive or to non-cursive scripts. The text in natural images and that in printed or scanned documents do not share any commonalities. That is why these approaches drastically failed on scene text recognition systems.

Numerous ideas have been proposed to address the complexities involved in scene text recognition. However, it may be categorized as a problem of the OCR field that contemplates different approaches to recognize scene text. The accuracies reported on various OCR techniques hone scene text images in the presence of non-text patterns. However, scene text recognition is labelled as a specialized problem of OCR, but there are distinct issues relevant to scene text recognition in comparison to typical OCR systems. One of the prominent issues in scene text is localization of a text from natural scene images. There are numerous techniques presented to address text localization as explained in [21–24] and text classification as examined in [25–28], respectively. This survey summarizes recent techniques designed for both phases, i.e., feature extraction and classification in scene text recognition. Active research contributions on scene text have been witnessed during past few years. Primarily, the proposed techniques have been applied to localization and recognition of Latin text from natural images [9,10,14]. The cursive script postulates more of a challenge to recognize text from the acquired image. As mentioned earlier, Arabic script is cursive in nature due to the inherent variability of single and joining characters [4,29].

Figure 3 shows the Arabic and Latin camera-captured scene text.

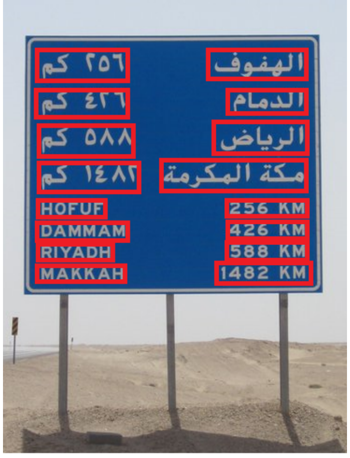


Figure 3. Cursive and non-cursive captured text image.

In recent years, a few research works have been reported [30–32] on Arabic text recognition especially in video images, but there is no standard dataset available for the aforementioned purpose. The work reporting on Arabic scene text is relevant to the images clipped from videos, as explained in [32–34]. Today, due to advanced media in every country, there are numerous international, national and regional news telecasts by each country. Thus, video text is easier to get in comparison to the proposed camera-captured scene text images. The text represented in a news video can be categorized into two types, i.e., artificial text and scene text. The text that is artificially overlaid on a video image is treated as prior, whereas the text image taken from camera during video capturing is classified as later. It is obvious that OCR will not directly process the video image because the nature of OCR is more towards processing clean document images taken at standard resolution and in specific settings. The video images often have colour blending, blur, low resolution and complicated backgrounds in the presence of different objects.

This paper presents a comprehensive literature survey on cursive scripts, especially Arabic scene text recognition. The most influential and convincing work proposed in recent years is summarized in this paper. The state-of-the-art techniques for text detection have mainly been summarized by [35]. Another survey was compiled by [23] to discuss detection and recognition techniques, but it lacked recent state-of-the-art techniques. A very persuasive survey paper was written by [23]. They mostly discussed the latest techniques that were established in recent years for text detection and recognition, but with respect to Latin. Furthermore, they also discussed future trends.

The motivation behind the contribution of this survey is as follows:

1. Scattered works have been seen on cursive scene text recognition, especially in Arabic script. The aim is to provide detailed knowledge to researchers about the current status of work for the purpose of discussing the future possibilities of research in Arabic scene text analysis.
2. This survey provides a substantial contribution from the researcher's point of view to address the inherent complexities as explained in [4,5,8,30] and proposes an idea of how to overcome them.
3. The details about state-of-the-art techniques are also provided, which exhibit the research on cursive scene text recognition and may assist Arabic text recognition in natural images.
4. The details about available Arabic scene text datasets are also provided, which may guide the researcher about what level the research has been done in Arabic scripts and what are the hindrances during the process.
5. This paper provides new insights to researcher and gives us an idea where Arabic or Arabic-like (such as Farsi, Urdu) languages stand in scene text recognition field.

The presented survey is organized into different sections. The complexities in scene text recognition are summarized in Section 2. This section further summarizes the details about the solutions presented for text localization and classification. In addition, it also provides detail about available Arabic datasets. The advances of the deep learning network in text analysis are presented in detail in Section 3. Section 4 narrates discussion, while future directions are provided in Section 5. This survey is concluded in Section 6.

2. Complexities Involved in the Scene Text Recognition Process and Its Solutions

Text detection and recognition are contemplated as subtle tasks in scene text analysis. The three basic challenges identified by [35] are diversity in text, background complexity and embedded noise in text created by interference. The scene text may be overwhelmed by variations with respect to font size, colour, noise and inconsistent background. It becomes a challenge to recognize the text in the presence of such tricky impediments. The phases involved in scene text recognition are presented in Figure 4.

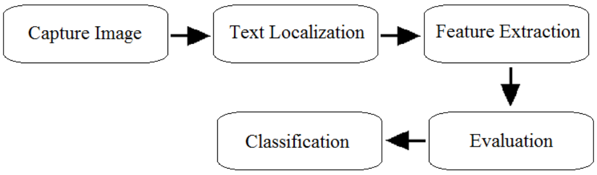


Figure 4. Scene Text Recognition (STR) phases.

To counter the implicit problems associated with scene text, there exist various proposed techniques that address the inherent challenges faced for text appearing in natural images. These techniques achieved significant results, but all of them were applied to Latin script, as reported in [9,14]. The following sub-sections elaborate each phase of the text recognition process in detail.

2.1. Methods Designed for Text Localization

Text detection or localization is regarded as an important part of information extraction systems. During recent years, novel approaches for text detection have been proposed by the pattern recognition, document image analysis and computer vision research communities [36]. In most of the surveyed work, it is observed that the script identification in multi-script scene text images is presented as a prominent focal point of current researcher's interest. The text detection and localization algorithms are not designed for a specific language. Instead, this process is the same for any type of language, but the recognition techniques may vary depending on the nature of the script's characteristics and complexity. Numerous methods have been reported in the past few years that describe the correct localization of text in natural images. Some of them have been explained in the following paragraphs with their reference.

Conditional Random Fields (CRF): This is a probabilistic method that predicts the sequence of missing information between the previous and current sequence of content appearing in order to make an exact or approximate label. One such work was presented by [37]: they proposed a hybrid approach for text localization, which was based on CRF. Their proposed method consisted of three stages, i.e., pre-processing, connected component analysis and minimum classification error. The connected component technique may lead to the problem of inaccurate localization of text; hence, CRF is employed to predict the accurate label.

Scale-based region growing: Text detection based on region growing was proposed by [11]. The process of region growing starts from the keypoints detected by the Scale-Invariant Feature Transformation (SIFT) algorithm [38]. By the SIFT algorithm, the keypoints of a given image were extracted, which is treated as a feature regardless of extracting the keypoints that appear in a text or a non-text region. Their algorithm defined the region's range and summed up all keypoints that fall in the range specified for the region; in this way, the region gradually grows larger. The extracted text blocks map geometrically, which sometimes includes background lines as a candidate region, which might decrease the OCR accuracy. To address this problem, they proposed a fast text block division algorithm, the detail of which can be seen in their manuscript [11]. Figure 5 shows the scale-based region growing with the identification of keypoints.

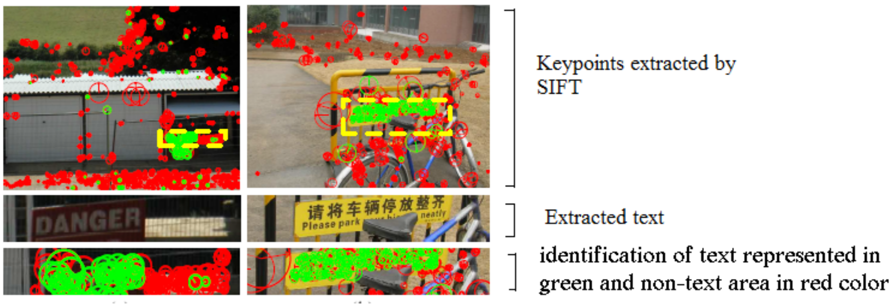

Keypoints extracted by SIFT
Extracted text
identification of text represented in green and non-text area in red color

Figure 5. Scale-based region growing with the identification of keypoints and marking of a text.

Oriented stroke detection: The text stroke orientation may also play a vital role in scene text detection, as presented by [14]. The assumption lies in the fact that every character is represented by its stroke information. They measured the orientation of the character stroke. For instance, character "*A*" is represented by two stroke directions, one that is 60 degrees on the left and the other 180 degree on right, and both are joined on the top. One stroke makes zero degrees, which joins the left and right stroke from the middle. The gradient projection "*G*" of every character is modelled and noted down direction α with a scale "*s*", as described in their paper as follows,

$$G_{\alpha,s} = \frac{\beta R_{\alpha} S_s I}{\beta} \tag{1}$$

R_α is represented as a rotation matrix of an angle α; whereas, S_s is labelled as the scaling matrix of a scale s. The detail about their proposed method can be seen in the manuscript [14]. The candidate region is detected, and it grows by detecting at least a single stroke in an image I.

Another work based on strokelets to detect multilingual text from scene images was proposed by [39]. The stroke information of every character has been taken into account and compared with the rest of the text for character recognition. The overall process was to find the centre of a candidate character by seeking maxima in the Hough transform using mean shift. The weighted average was used to determine the candidate cluster. A template of a window size of 5 × 7 was suggested to store the stroke information of each character. They used 63 classes, which were divided into 10 Arabic numerals and 52 English letters and one special class to handle the appearance of invalid characters. The evaluation was performed on two publicly-available datasets, i.e., dataset defined in International Conference of Document Analysis and Recognition dataset ICDAR2003 and Street View Text (SVT) [40].

Text detector based on the Laplacian and Sobel filter: The orientation of text in video images portrays a challenge for researchers due to the constraint of the size and orientation of a text image. To address the problems and challenges in multi-orientation in video text images, Shivakumara et al. [34] proposed a text detection approach. They applied the product of the Laplacian and Sobel filter for enhancement of pixel values. The discontinuities in an image gave them a strong indication about the presence of text. They used a 3 × 3 mask to get the enhanced text pixels. The Laplacian is a second order derivative and is used to detect dissimilarities in four directions, i.e., horizontal, vertical, up left and up right. In this way, information of low or high contrast images was enhanced. However, the results they found were noisy. For the purpose of handling this noise, they used the product of the Laplacian and Sobel filter. The Sobel filter is a first derivative, and it produces a fine detail at discontinuities in the horizontal and vertical directions. They first individually applied both techniques and later combined the result as a product of both techniques.

Connected component analysis: This is a labelling technique that scans an image and groups its candidate pixels in a way that all connected pixels make a component that shares similar intensity values and some how their pixels get connected to each other. This technique is applied to natural images or to video text images for determination of text area. The video image often has a complex background, cluttered text and a jerking image. In this situation, it becomes very difficult for any type of method to produce encouraging results. The text localization using the stroke filter in video images was proposed by [41]. They designed a stroke filter in such a way that uses the properties of local feature analysis and global constraints. They employed connected component analysis for text localization in text regions. They applied the stroke filter on the source image and later analysed it by the connected component analysis technique. They compared the stroke filter with Cannoy, Gabor, Haar and ratio filters and came to the conclusion that the stroke filter was best among them.

Co-occurrence of the histogram of oriented gradients: Another cursive scene text detection and recognition work was presented by [42]. They proposed scene character recognition using the co-occurrence of the histogram having oriented gradients. The information about the oriented gradient of the image pixel is a point of interest, while ignoring its neighbouring pixel's values. The relevant pixels were selected based on the maximum offset represented in the horizontal and vertical directions. In this way, pixels made an orientation pair with reference to the central pixel, and covariance was accumulated at the end. The elaborated description of said approach may be found in their paper. They evaluated their technique on Latin, Bengali and Chinese script. They captured 260 Bengali scene text images. On the other hand, they collected 487 Chinese scene text images, and out of them, 3419 were used for training and 2763 for testing purposes. They also evaluated their proposed method on a Chinese dataset named ChiPhoto, which consisted of 343 scene text images. The ChiPhoto dataset includes classes that are not included in the Pan_Chinese_Character dataset. All Chinese images were normalized to x-height 13–365 pixels, while width was adjusted from 11–325 pixels, as explained in their research. They also mentioned their evaluated results on the English dataset. Their best result was reported on convolutional Co-HoG (Histogram of Gradients)

Maximally-Stable Extremal Regions (MSER): In computer vision, this is a well-known method for blob detection. It was first introduced by Pajdla et al. [43]. The correspondence between two image elements with two different view points was considered. The assumption was to extract a comprehensive number of image elements that contribute to matching the baseline, which helps with the detection of an object in an image. This algorithm has been applied to detect text candidates in various state-of-the-art applications [44–47]. A multilingual text detection through MSER was proposed by [48]. They used MSER to detect text from the scene image. The input for the MSER algorithm is a greyscale image I_g, and the output is I_t, where $t = 0$–255. They first binarized the image with a threshold t. Every pixel was evaluated and changed into a black or a white area, where zero means completely black and 255 completely white. The white area in the image is called the extremal region. To detect extremal regions, the rest of the pixel area should be the same. The threshold T was applied to take the exact number of interested regions. By applying threshold T over the image, they obtained successive regions that were not impacted by the overall process. Such regions are said to be Maximally-Stable Extremal Regions (MSER).

In another manuscript, the candidate regions were detected by MSER [22]. As the nature of MSER suggests, the individual characters are not detected correctly by it because most of the times, it is used to find the region of interest. Hence, it has been proven to be suitable to find the number of characters or words. In general, MSERs in an image is categorized into three classes. The first class corresponds to individual characters, while the second class may have an arbitrary number of characters, whereas the third class may contain all non-textual background content. They estimated characters on the basis of character strokes. The character area was expanded by calculating the distance transform map, which depends on the basis of the calculated binary mask. Here, the pixels played an important role, and local maximum distance was considered for extremal region detection. For the purpose of calculating character stroke area A_s [22], the following equation is used,

$$A_s = 2\sum_{i\in S} d_i \tag{2}$$

where S is the stroke and d_i is the distance of pixel i to the boundary. Their proposed estimation was taken as correct only for the strokes that have an odd width, while it became inaccurate on even widths. The boundaries of stroke pixels were not connected to each other because of this noise. This has been compensated by introducing weight w_i as follows,

$$A_s = 2\sum_{i\in S} w_i d_i, \quad w_i = \frac{3}{|N_i|} \tag{3}$$

where N_i denotes the number of stroke pixels in the 3×3 matrix.

Another very interesting manuscript describing MSER was proposed by [47]. They exploited the characteristics of MSER to detect text from given images. For efficient processing, they predicted extremal regions in an image and took computational complexity into consideration. The search space was also limited by linear timing.

After examining numerous manuscripts describing the use of MSER, it is observed that MSER is a very important and suitable technique to detect character candidates from a scene image. It is also considered as an invariant to affine transformation. By keeping its ability to search the point of interest in the provided area, it can produce good results with low quality images. The implementation complexity is $O(nlog(log(n)))$, where n is the number of pixels in an image. Although this algorithm is more suitable for text detection, in some situations, it might detect false positives, which can further be investigated by applying various checks for the purpose of eliminating regions that are not of interest from a given image.

Toggle mapping: The work presented by Fabrizio et al. [9] is based on toggle mapping method to segment the text from natural scene images. It is defined as a morphological operator first introduced by [49]. Toggle mapping maps the function f onto a set of functions. It is primarily used for contrast enhancement and noise reduction. As explained in [9], the proposed method is used to segment the greyscale image I into two sets of functions $h1$ and $h2$. The morphological erosion of I was done by $h1$, and morphological dilatation of I was performed by $h2$. Hence, their proposed method can detect the boundary, but introduces salt and pepper noise on homogeneous regions. In order to control the noise appearing in homogeneous regions, it is essential to take a value that represents the homogeneous region. Therefore, three parameters were taken into consideration, i.e., $h1$, $h2$ and minimal contrast C_{min}. They evaluated their proposed method on 501 readable characters and compared their results with the Ultimate opening, Sauvola and Niblack filters. They found a high percentage, i.e., 74.85% of correctly-segmented characters in their proposed method for text segmentation.

Graph-cut method for scene text localization: The scene text localization method based on the graph cut approach was presented by [24]. The edges of an image are first extracted through the local maximum difference filter, and the,n the image is clustered based on colour information. The candidate regions contact the text that was identified by combining geometric structures. The spatial information of scene text in a skeleton image was generated by extracting the edges. The characters were realized by applying heuristic rules in addition to connected components. At the end, they applied the graph cut approach for the purpose to identify and localize text lines correctly.

They performed comprehensive evaluation experiments on four datasets (i.e., ICDAR2003, ICDAR2011, MSRA-TD500 and the Street View Text (SVT)) for the purpose of validating their proposed method. They concluded that their approach produced state-of-the-art results on diverse fonts, sizes, and colours in different languages regardless of the impact caused by illumination.

2.2. Methods Designed for Feature Extraction

In the context of machine learning and pattern recognition applications, features are considered as the backbone of any recognition system. In this sub-section, the importance of feature extraction techniques by providing the analysis of recently-proposed methods reporting good accuracy is presented.

The feature values derived from the raw or initial set of given data intend discriminative and non-redundant data. This discriminative and non-redundant data facilitate the process of further classification and learning steps. There are numerous feature extraction methods proposed by various researchers designed specifically for scene text images. The detail about prominent methods is depicted in the following paragraphs.

Global sampling: As described in [35], the scene text character's recognition performance is evaluated by considering the comparison of different sampling methods, i.e., local and global sampling, feature descriptors [16,50–52], dictionary sizes, coding and pooling schemes [53] and Support Vector Machine (SVM) kernels [15]. To obtain the features from local sampling as mentioned in [35], the keypoint detection, compute local descriptors, build dictionary of visual words and feature pooling and coding to get the histogram of visual words are important to investigate due to their discriminative nature. They computed the descriptors from a character that is patched by global sampling without considering keypoints, local descriptors, coding and pooling. The features they found by following their process are regarded as distinct features that are ready for classification.

Multiscale histogram gradient descriptors: The multiscale histogram of oriented gradient descriptors as a feature vector was reported by [10]. They included features at multiple scales in a column-wise manner of the HOG descriptor and evaluated the performance on Latin scene text images having variation of characters. The oriented gradients were calculated using the derivative of Gaussian filters. Then, the total strength of each variation was summed up, and this process continued on each block in an image. Each histogram with respect to each block was normalized so as to sum up the total variation across all orientations. In this way, they made a single descriptor for each image. They evaluated their proposed technique on two commonly-used datasets, i.e., Chars74K [54] and ICDAR03-CH [55]. Chars74k contains 62 classes, which include number, upper and lower case characters. The other dataset they used was ICDAR03-CH, which is the character recognition dataset that was presented at the robust reading competition in ICDAR2003. This dataset is similar to Chars74k. In addition, it included the images of punctuation symbols. They split the Chars74k dataset into Chars74k-5 and Chars74k-15 training images per class. The ICDAR03-CH-5 dataset was used with four training images per class. The reported accuracy using multiscale HOG was 50%, 60% and 49% and by HOG columns was 59%, 67% and 58% on Chars74k-5, Chars74k-15 and ICDAR03-CH-5, respectively.

Scale-Invariant Feature Transformation (SIFT): This is a main computer vision algorithm that is used to define the local features of an image. The local features that it detects are the keypoints that are not affected by image transformation. The extracted keypoints from the SIFT algorithm are depicted in Figure 6. The scale-invariant approach is applied to Arabic scene text recognition [31] with the combination of sparse coding [56] and spatial pyramid matching [57], as shown in Figure 7.

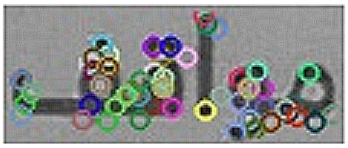


Figure 6. Scale-invariant feature transformation of Arabic text.

They extracted the local features from SIFT [38], which is considered as a very efficient technique that demonstrates and extracts the most relevant distinguished local features. In order to get more precise information of an image, the weighted linear super-position function is applied to extracted descriptors.

The input image was divided into sub-regions, then features relevant to a specific sub-region using different scales were modelled into the histogram. Later, they applied the pooling technique to summarize all the features representing the image. The evaluation was performed on two publicly-available datasets, i.e., Chars74 and ICDAR03, and reported 73.1% and 75.3% accuracy, respectively. They also proposed their own dataset named Arabic Scene Text Characters (STC) and evaluated the performance of their proposed system on it. Moreover, they reported 60.4% character recognition accuracy on their proposed dataset.

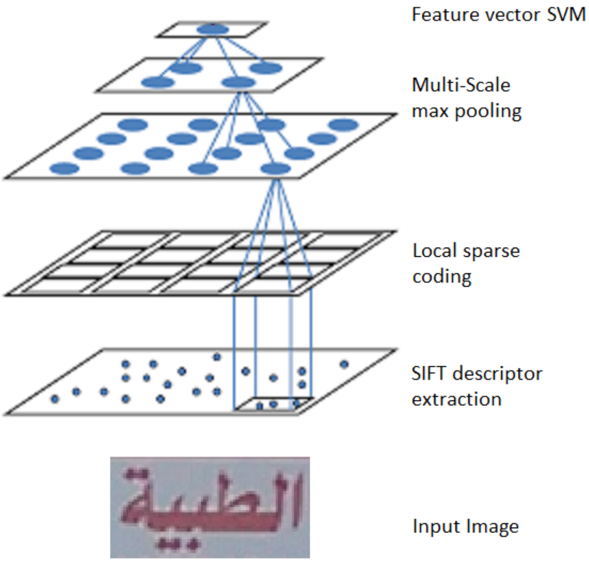


Figure 7. Character classification through SIFT features and sparse coding.

Another paper that represents Chinese handwritten character recognition by considering SIFT descriptors was proposed by [58]. They modified the SIFT descriptor according to the characteristics of Chinese characters. The pre-processing was performed by passing each image through linear normalization and then performing elastic meshing [59] for the purpose of rectifying the invariance of the same characters written by various individuals. Moreover, they also extracted Gabor and gradient features of an image. Every extracted feature vector was compressed to 256 dimensions by Linear Discriminative Analysis (LDA). They performed experiments on different window sizes with various dimensions. The discussion about their detailed experiments can be found in their manuscript.

A very interesting work on local features extraction based on template images was proposed by [60]. Their motive was to read the text in complex images in the presence of built-in noise associated with it. They proposed a new method for building the template in the absence of influential noise. After performing normalization, enhancement and binarization, they extracted scale-invariant features from the template image and from the complex image, as well. If some features were missing or not recognized, then their proposed geometrical verification algorithm was applied to correct the error. They evaluated their proposed technique on more than 200, 000 images having three scripts, i.e., Chinese, Japanese and Korean script. On single character recognition, they obtained 94.1% accuracy, while on multiple character recognition, they obtained 89.6% accuracy.

Hybrid features: In [48], the hybrid feature extraction approach was proposed by determining the stroke width, area, aspect ratio, perimeter, number and area of holes as a feature associated with each given text image. These features were examined and further passed to the classifier. The description about each feature is detailed in their manuscript. Their proposed method was evaluated on various benchmark datasets like Latin and multilingual scripts. The dataset, named as MSRA-TD500 , ICDAR2011, ICDAR2013, except the ICDAR2011 dataset, and other datasets contain multilingual texts including Hindi and Arabic text, as well. Furthermore, they also evaluated the performance of the proposed algorithm on their collected data samples. L. Neumann et al. [14] used eight different hybrid features of the text detected by MSER. The identified features were character width, character surface, aspect ratio, stroke width, character height, character colour, vertical distance bottom line and MSER margin. They passed these features to the classifier to train the network. The two publicly-available datasets ICDAR2003 and Char74K dataset were evaluated. The reported accuracy was 74% on the Char74K dataset, while on ICDAR2003, they achieved 62% accuracy. Table 1 summarizes the detail about feature extraction approaches that have recently been proposed for cursive and non-cursive scripts.

Table 1. Feature extraction approaches of cursive and non-cursive scene text.

Study	Feature Extraction Approach	Script	Dataset Used
Newell et al. [10]	Histogram of Oriented Gradients (HOG)	Latin	Char74k , ICDAR03-CH
Neumann et al. [14]	Stroke orientation	Latin	ICDAR2011
Tounsi et al. [31]	SIFT	Arabic, Latin	ARASTEC
Yi et al. [35]	Global and local HOG	Latin	Char74k , ICDAR2003
Tian et al. [42]	HOG	Chinese and Bengali	IIIT-5k word, Pan Chinese, ISI Bengali Characters
Wu et al. [59]	Minimum Euclidean distance, SIFT	Chinese	ETL9B
Zheng et al. [60]	SIFT	Chinese, Japanese, Korean	Datasets A, B, C, D, E (own compiled)
Campos et al. [54]	Geometric blur, shape context, SIFT, Patches, spin images , Maximum Response (MR8)	Latin, Kannada	Own compiled
Gomez et al. [61]	Convolutional neural network and K-means	Multilingual	CVSI, MLe2e dataset
Mao et al. [11]	SIFT	Latin, multilingual	ICDAR, SWG, MSRG

The IIIT-5k as mentioned in Table 1 is a dataset compiled by International Institutue of Information Technology (IIIT) whereas, Indian Statistical Institute (ISI) compiled ISI-Bengali characters dataset. The ETL Character database was prepared by the Electrotechnical Laboratory (ETL) in Japan.

2.3. Classification Techniques for Scene Text Recognition

Classification refers to the statistical analysis of training observations under supervised and unsupervised learning. The classifiers analyse the numerical properties of a given image. These numerical properties are the distinctive features that represent the image in question. There are numerous state-of-the-art classifiers that have been proposed during recent years, but most of them depict the process of learning through unsupervised methods [15,62]. The summarized version of a few proposed classification techniques with their references are elaborated as follows.

Nearest neighbour classifier: The nearest neighbour classifier is categorized as a non-parametric method used to train the given sample based on the closet trained neighbour in the feature vector. An interesting work representing nearest neighbour as the feature vector was proposed by [14]. The candidate regions that have been extracted by oriented stroke information are labelled by Unicode. In their experiments, they took the extracted region as black in colour with a white background. They considered 62 character classes written in 90 font styles. In total, they had 5580 training samples.

In another manuscript, the nearest neighbour classifier was used by [63] to train and classify the given pattern. For the purpose of obtaining the maximum performance, they set the value of K to 11. They evaluated their proposed technique on the ICDAR2011 Robust Reading Competition dataset [2] by using an evaluation protocol of ICDAR2011 [64]. The results were reported as recall of 66.4%, while the precision was measured as 79.3%. They concluded that their recall results were better than the winner of the ICDAR2011 Robust Reading Competition, which reported 62%, and Shi's method, which reported 63% [65].

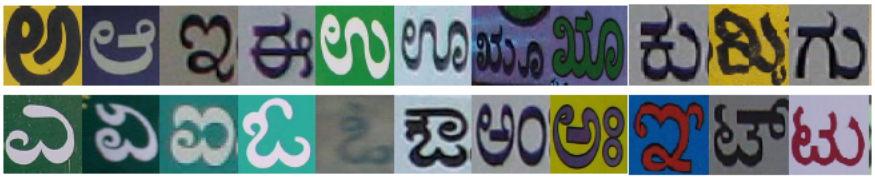
Neural network as a classifier: The artificial neural network is considered as a reliable classifier, which is inspired by the human way of learning things. A recent and novel work on isolated Arabic scene character recognition using a deep learning classifier was proposed by [13]. They used a convolutional neural network and trained the classifier on different character orientations. They further formulated the training on 3 × 3 and 5 × 5 filter sizes by keeping the stride values of one and two. Moreover, various learning rates have been applied for the purpose to get maximum accuracy. They reported a 0.15% error rate on the recognition of isolated characters.

The work presented by Mao et al. [11] is scale-based region growing technique to detect text in an image, while the neural network was proposed to learn the pattern. They decomposed the input into 128 dimensions, which is further passed on to a hidden layer network of 40 in size. They trained 67% of the dataset samples, while the remaining 33% were used to test the trained network. They evaluated their technique on ICDAR2003, ICDAR2005, the Stroke Width Transformation (SWT) dataset and their proposed Multilingual scale based region growing (MSRG) dataset. The SIFT keypoints were extracted and reported to have 81% accuracy on ICDAR, 87.90% on SWT and 86.46% on MSRG, respectively.

The latest work on Latin or English using Convolutional neural Networks (ConvNets) was recently proposed by [61]. They proposed a multi-stage approach for script identification. At first, the text image was normalized by an x-height of 64 pixels while maintaining its aspect ratio, then they applied the sliding window approach to extract 32 × 32 image patches, which they called stroke parts. Later, each part was presented to ConvNets to get feature values. Furthermore, L.G Bigorda et al. [61] explained that each text line is represented as a stroke part's descriptor. They designed the same structure and followed the same pre-processing steps as explained in [62]. They employed the K-means algorithm on extracted patches to learn the convolutional kernels of ConvNets. Instead of extracting convolutional features from a single image with a single vector as performed by [62], a set of convolutional feature vectors was extracted from parts of the image. If applying ConvNets on extracted image patches, it cannot provide stable and good convolutional features because of the inherent complexity attached to cursive scripts. Hence, they further suggested an idea of merging the Naive Bayes Nearest Neighbour (NBNN) classifier [66] with their proposed solution. The notion behind the said classifier is to compute the image directly without defining intermediate quantized descriptors. Thus, all extracted image patches were given to NBNN to compute the maximum difference of all provided templates that exist in other classes except for its own class. They reported the number of experimental variations on two publicly-available datasets and their own prepared dataset named Video Script Identification Competition (CVSI-2015) [33], the ICDAR 2013 dataset and their proposed dataset named MultiLingual end to end (MLe2e) , which is comprised of 711 multilingual scene text images. Among various experiments as described in their manuscript, they also performed cross-validation of their trained network on the above-mentioned datasets. The trained network on their proposed dataset MLe2e obtained higher accuracy on given test images of CVSI and ICDAR 2013, which was 70.22% and 94.70%, while on MLe2e, they obtained 91.67% accuracy.

Hybrid classifier: T. E. de Campos et al. [54] proposed three classification schemes, i.e., the Nearest Neighbour (NN) classifier, SVM and Multiple Kernel Learning (MKL). They performed detailed experiments and reported results on their collected dataset. The collected samples contained Latin and Kannada text. There were 657 numbers of classes for Kannada text with few variations. The determination of accurate text is still a cumbersome task in text detection in natural image. Thus, they manually localized the text. Moreover, they gathered characters generated through synthetic means.

The proposed technique was evaluated on 1922 natural images, 3410 handwritten English characters and 16425 Kannada characters written by 55 and 25 volunteers, respectively, as presented in Figure 8. They considered six different types of local features such as shape contexts [17], geometric blur [18], scale-invariant feature transform [67], spin image [68], maximum response of filters [69] and patch descriptor [70]. The detail about each feature can be seen in [54]. They also evaluated their proposed method by ABBY FineReader; furthermore, the reported results were obtained in the ICDAR Robust Reading Competition 2003 and 2004, as well The detailed experimentation was performed, and the obtained results on each of the features using three different classifiers on Latin and Kannada script were obtained as mentioned in Tables 1 and 2, while Figure 8 represents the Kannada text sample.



ಮಹಾನಿಘಂಟಿನ ಎಂಟು ಸಂಪುಟಗಳನ್ನು ಅದು ಪ್ರಕಟವಾದ ೧೫ ವರ್ಷಗಳ ನಂತರ ಮತ್ತಷ್ಟು ಸುಂದರವಾಗಿ
ಮರುಮುದ್ರಣಗೊಳಿಸಿ ಕನ್ನಡ ಕುಲಕೋಟಿಗೆ ಅರ್ಪಿಸಲು ನನಗೆ ಸಂತೋಷವೆನಿಸುತ್ತದೆ. ಕನ್ನಡ ಶಾಸ್ತ್ರೀಯ
ಭಾಷೆಯ ಸ್ಥಾನಮಾನವನ್ನು ಪಡೆದುಕೊಂಡಿರುವುದು ಅದರ ಸಂಪನ್ನತೆ ಮತ್ತು ಅರ್ಹತೆಗಳಿಂದಾಗಿಯೇ

Figure 8. Kannada cursive script including printed and scene text samples.

Support Vector Machine (SVM): This is another machine learning approach that is used to classify the data by performing regression analysis using supervised learning methods. In a survey paper presented by Yi et al. [35], the work on SVM is compiled and depicted good results on global sampling. The Char74K and ICDAR2003 datasets were used to evaluate their proposed methodology. All experimentation was performed by the SVM classifier [71]. The reported accuracy was 62% on the CHARS74K-15 dataset and 76% on the ICDAR2003 CH dataset using global HOG as a feature vector, while on local HOG, the reported accuracies were 58% and 75% on the above-mentioned datasets, respectively. The work proposed by Neumann et al. [72] also trained the SVM classifier on their proposed dataset. They evaluated the performance on the Char74K dataset by following the protocol defined by Campos et.al [54]. The classifier was trained on 7705 annotated characters extracted from 636 images. They also created a language model with 1000 frequently-used English words. They evaluated the character recognition by considering three situations, i.e., a character has been localized and recognized correctly, so the given character is matched; the second situation is when a character has been localized correctly, but not recognized, which led to the problem of the mismatched case. In other words, when the character was not correctly localized, this meant that the character was not found. They compared their results with Campos et al. [54] and found that their technique produced better results, which was 71.6% in comparison to 54.3%.

K-means clustering: This is one of the most prominent and simple unsupervised learning algorithms that classifies the given data through a certain number of clusters. One such work was presented by [62]. They proposed an unsupervised learning algorithm that generates features that were used for classification. They guesstimated the variant of K-means clustering and compared their yielded results with other methods. The input image was normalized by compressing the image into a

32 × 32 image size and then collecting 8 × 8 grayscale patches of images, then applying the statistical preprocessing whitening technique as proposed by [73] to the yielded patches, which helped them to make another dataset. They took a number of patches for the purpose of obtaining the vector of pixels, which was normalized with respect to contrast and brightness, and put them in the dictionary. As they applied the K-means variant of clustering, their classifier learned the set of normalized vectors using the inner product as the similarity measure. They evaluated their ICDAR 2003 dataset using linear SVM and reported character recognition accuracy on three different dataset classes, as further explained in their paper. The highest accuracy they reported was 85.5%.

Bayesian classifier: The Bayesian classifier is a probabilistic classifier based on the Bayes theorem, which helps to predict the independent assumption about the features. One such work was presented by [34]. The text candidates were obtained by intersecting the output of the Bayesian classifier and the canny edge map of input image. They assumed that the orientation and size of a font does not have any impact on recognition due to the proposed Bayesian classifier. The connected component method was used to make a bounding box of a text. They evaluated the performance of their proposed technique on their collected dataset named Hua [74] and ICDAR 2003 [55].

Minimum Euclidean distance classifier: This is suitable to compute unknown patterns and predict the decision based on the smallest distance. The proposed work by Zhang et al. [58] presented the minimum Euclidean classifier to learn the unknown patterns of Chinese handwritten script recognition. The experimental data contained 3755 frequently-used simplified Chinese characters, which were normalized into 64 × 64 pixels. This dataset was collected by Beijing University of Posts and Telecommunications. The reported results were produced on 6, 7, 8 and 9 subregions. The best result was computed by fixing the sub region on seven and obtained 94.80% accuracy. In another experiment, they evaluated their proposed amended SIFT feature's (Char-SIFT) performance compared to the standard SIFT technique. Their result showed good accuracy as 94.66% on linear discriminative analysis; whereas, on each feature, i.e., Gabor, gradient and Char-SIFT, they reported accuracies of 96.58%, 97.52% and 97.86%, respectively.

Fully-Convolutional Networks (FCN): Another very interesting paper on text detection and recognition was presented by [75]. The fully-convolutional network classifier was proposed and trained on multi-oriented Chinese text. They considered the local and global aspect of a given text for correct text localization. They first trained the model to predict the salient features of a given text image, then the hypothesis for the text line was estimated by combining salient features and character components. The false hypothesis was removed by FCN after computing the centroid. They reported very good results on three text detection benchmarks, i.e., MSRA-TD500, ICDAR2015 and ICDAR2013.

Table 2. Classification approaches applied to cursive and non-cursive scene text.

Study	Classifier Techniques	Script	Database
Newell et al. [10]	Not Reported	Latin	Char74k, ICDAR2003-Characters
Tounsi et al. and Yi et al. [31,35]	SVM	Arabic, Latin	ARASTEC, Char-74k, ICDAR2003
Neumann et al. [14]	Nearest neighbour	Latin	ICDAR2011
Gomez et al. [61]	Convolutional neural network and K-means	multilingual	CVSI ′ MLe2e dataset
Campos et al. [54]	Nearest neighbour, SVM and kernel learning	Latin, Kannada	Own compiled
Mao et al. [11]	Neural network	Latin, multilingual	ICDAR, SWT, MSRG
Wu et al. [59]	Minimum Euclidean distance classifier	Chinese	ETL9B
Zheng et al. [60]	NR	Chinese, Japanese, Korean	Datasets A, B, C, D, E (own compiled)
Zhang et al. [75]	Fully-convolutional network	Chinese	MSRA-TD500, ICDAR2015 and ICDAR2013

2.4. Arabic Scene Text Datasets

The dataset plays a very dominant role in investigating the potential of numerous classifiers. Various efforts have been reported for capturing and preparing the datasets for Arabic text in natural images in the recent past. Some articles presented a survey on available open access datasets and tools specifically designed for Arabic text detection and recognition in video frames captured by news channels [76]. The benchmark dataset for Arabic scene text still requires more effort to standardized the research as far as Arabic scene text analysis is concerned. The description of a few available datasets is revealed as follows. In addition to available datasets, this survey also describes the details of our collected data samples.

ARASTEC (ARAbic Scene TExt Characters) 2015: A recently-compiled dataset was prepared by [31]. ARASTEC is an abbreviation of ARAbic Scene TExt Characters. They captured 260 natural images containing Arabic text. Images were taken from sign boards, hoardings, and advertisements. They manually segmented characters from images and obtained 100 classes depending on the position of a character in a word depicting 28 Arabic characters. They obtained 30–40 variations of each class. The sample Arabic text image of the ARASTEC2015 dataset is shown in Figure 9.

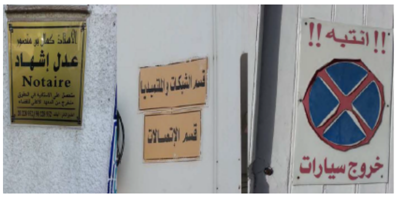


Figure 9. Sample image from the ARASTEC 2015 dataset.

ALIF: This is considered one of the first Arabic embedded text recognition datasets, proposed by [77]. The gathered samples were taken from various Arabic TV broadcasts, e.g., Al Jazeera, Al Arabiya, France 24 Arabic, BBC Arabic and Al-Hiwar Arabic. They localized Arabic text from 64 recorded videos. There is a wide variety of text specification like colours, styles and font size. In addition, the text visibility is also impacted by acquisition conditions, e.g., contrast, luminosity and background colour, which make the dataset challenging to apply to evaluate state-of-the-art techniques. The ALIF dataset contains 89,819 characters, 52,410 paws and 18,041 words. The data that were collected had more than 20 fonts. Their proposed network was trained on 4152 text images, which covered a wide variability of acquired text. The potential of the trained network was evaluated on three different test sets. The first test included 900 images, which were selected from the same channels used during training. The second test set was applied in the same setting as the previous one, but with 400 additional images. The third test set had large variations of text with respect to font and size. It had 1022 text images in total. The complexity associated with camera-captured text images is different in comparison with printed text, as shown in Figure 10.

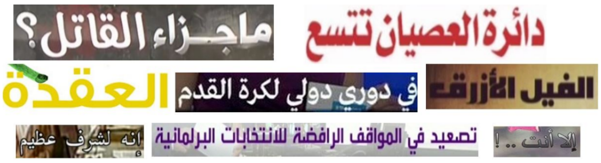


Figure 10. Embedded Arabic text of the ALIF dataset.

APTI (Arabic Printed Text Image): The APTI database was introduced by [78]. They generated Arabic data synthetically with font variations, different sizes and different styles like bold/italic. The samples were also decomposed into ligatures. The lexicon was divided into 113,284 words written in 10 different Arabic font styles by fixing 10 font sizes, but in four different font styles. Their dataset had 45,313,600 single-word images with more than 250 million characters.

In camera-captured text images, there is a need to deal with non-text captured objects, which are categorized as noise in an image that should be removed as a part of the pre-processing. Another challenging factor attached to camera-captured text images is illumination variation. The details from the given facts suggest that extensive pre-processing is required for camera-captured text image. As the APTI dataset was generated synthetically, this dataset can be treated as another category, because it was generated differently, as shown in Figure 11.

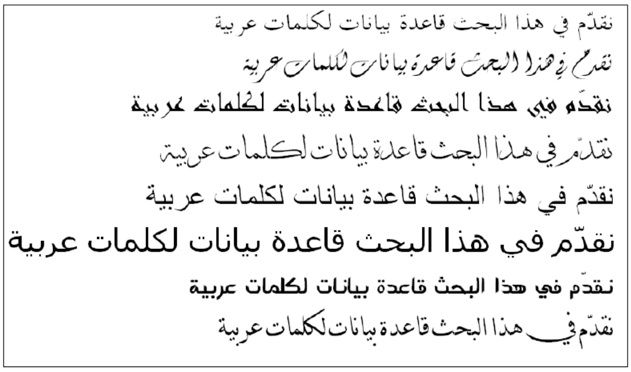

نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
نقدّم في هذا البحث قاعدة بيانات لكلمات عربية

Figure 11. Various fonts used to generate the Arabic Printed Text Image (APTI) dataset.

English-Arabic Scene Text Recognition (EASTR) dataset: The unavailability of an Arabic scene text dataset is a main hurdle in performing detailed research in this particular script as far as Arabic text recognition in natural images is concerned. In most Middle Eastern countries, the text appearing on signboards and hoardings is usually written English and Arabic, which prompted us to prepare a dataset for English in addition to Arabic, as shown in Figure 12.

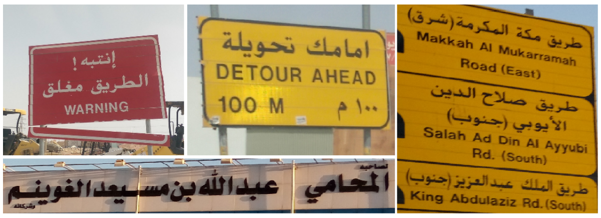

إنتبه !
الطريق مغلق
WARNING
امامك تحويلة
DETOUR AHEAD
100 M
١٠٠ م
طريق مكة المكرمة (شرق)
Makkah Al Mukarramah Road (East)
طريق صلاح الدين الأيوبي (جنوب)
Salah Ad Din Al Ayyubi Rd. (South)
طريق الملك عبدالعزيز (جنوب)
King Abdulaziz Rd. (South)
عبدالله بن مسيعد الغوينم
المحامي

Figure 12. Examples images of the English-Arabic Scene Text Recognition (EASTR)-42k dataset.

By considering an uncontrolled environment, we attempted to capture the maximum variations of Arabic text. The collected samples were segmented into English and Arabic text lines and numerals. EASTR-42K covers a huge variety of English and Arabic scene text appearing in uncontrolled environments. The details about the EASTR-42k collection are briefly elaborated in Table 3.

Table 3. EASTR-42K division based on complexity.

Language	Text Lines	Words	Characters
Arabic	8915	2593	12,000
English	2601	5172	7390

3. Advances of Deep Learning Network in Text Analysis

Intelligent applications assist humans in a wide variety of fields. Although they are not an ultimate replacement of human work, they are specialized in some aspect of human behaviour and produce results in a very efficient manner. Machine learning, which is considered as a sub-field of artificial intelligence, takes a leading role in exhibiting efficient behaviour. The machine learning systems are implemented with deep learning networks where the hidden layer network is deeply interconnected following a deep learning architecture. In recent years, machine learning has produced state-of-the-art results in Natural Language Processing (NLP) applications, especially in text analysis. The deep learning architecture produced impressive advancement in text analysis, as witnessed in [79–83]. Due to the densely-connected hidden layer, the deep learning network presented superior results on printed cursive text. As surveyed, the ConvNets and RNN-based LSTM networks are usually used to realize the deep learning architecture designed for cursive text analysis. The improvements can be suggested by adapting the LSTM architecture or by presenting a hybrid solution. This thesis proposes adapted ConvNets and the LSTM network by connecting the hidden layer neurons. The novel contribution is presented by incorporating the transfer learning experience with ConvNets on handwritten cursive script. Another contribution is to propose an adapted multidimensional LSTM network-based hierarchical subsampling approach on cursive scene text. The detail about the recent research presented using deep learning architecture is presented as follows.

3.1. Recurrent Neural Networks

Arabic script is represented in a contextual manner. The contextual learning is possible through RNN networks as each character in Arabic script depends on its previous character and predicts the next character at the current point in time. If context is eliminated from Arabic script, then recognition is not possible. Therefore, for cursive scene text analysis, the focus shifts from traditional backpropagation to context learning classifiers. Recent years' work as explained in [4–6,84] on Arabic-like script used the Recurrent Neural Network (RNN) approach for text classification. The RNN is suitable for problems where context is important to learn. The MDLSTM is considered as a connectionist approach, which mainly relies on the Multidimensional Recurrent Neural Network (MDRNN) and Long Short-Term Memory (LSTM) networks. The multidimensional LSTM follows the RNN approach to learn the sequences. All past sequences with respect to the current point in time are accumulated to predict the output character. The RNN provides an appropriate architecture for Arabic scene text recognition. The Arabic scene text image requires extensive pre-processing before applying the learning classifier. The pre-processing includes skew correction, removal of non-text elements from an image, conversion of the text image into standard format and feature extraction, which is a very crucial part machine learning tasks.

The deep learning architecture is implemented at the hidden layer of the LSTM network. As shown in Figure 13, the dotted lines marked at the hidden layer represent interconnection among memory blocks. The hidden layer can be adapted by inclusion of more than one layer, and the calculation at the hidden layer could be optimized by the interconnection of hidden layers at different levels, as adopted in the proposed thesis. The level of hidden layers and their interconnections among each other present the idea of the deeply interconnected network, which takes more time during gradient computation than usual, but exhibits good results in the learning of complex patterns, as experimented by [5,85]. Both works presented by Saeeda et al. rely on the deep MDLSTM architecture for learning the complex pattern of printed Urdu data. The deep MDLSTM network with its adapted architecture is followed in this proposed thesis, as it has not been extensively experimented on Arabic scene text recognition.

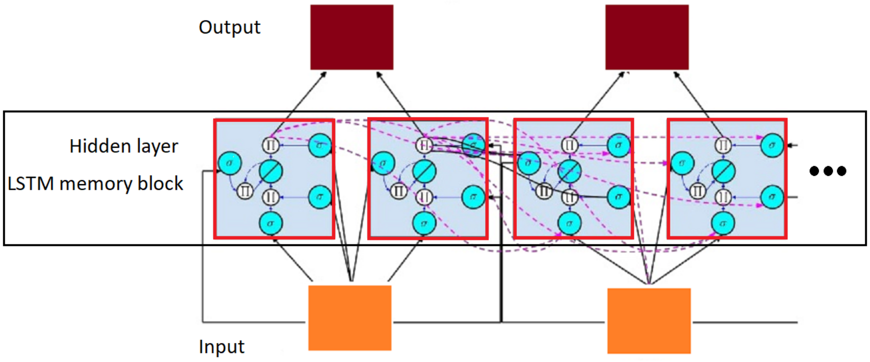


Figure 13. LSTM deep learning view.

3.2. Convolutional Neural Networks

Although ConvNets are suitable for feature extraction, they can be used as a learning classifier. The convolutional aspect of learning the convolved features by the shared weight architecture make ConvNets a deep learning architecture. ConvNets comprise a three-layer architecture where the hidden layer consist of convolutional layers, which are usually divided into activation function, pooling layers, fully-connected layers and normalized layers. The convolutional layers take data from the input layer, convolve them and delegate them to the next layer. The convolutional features are further processed or selected by the pooling strategy, i.e., the maximum pooling that takes the maximum value from each receptive field of neurons and at the earlier layer. The average pooling and minimum pooling can also be applied to convolved features. The receptive fields are square boxes containing neurons, as represented in Figure 14. The detailed feature extraction procedure using ConvNets is very much applicable to complex data like Arabic whether presented as a printed, scene text or as handwritten form.

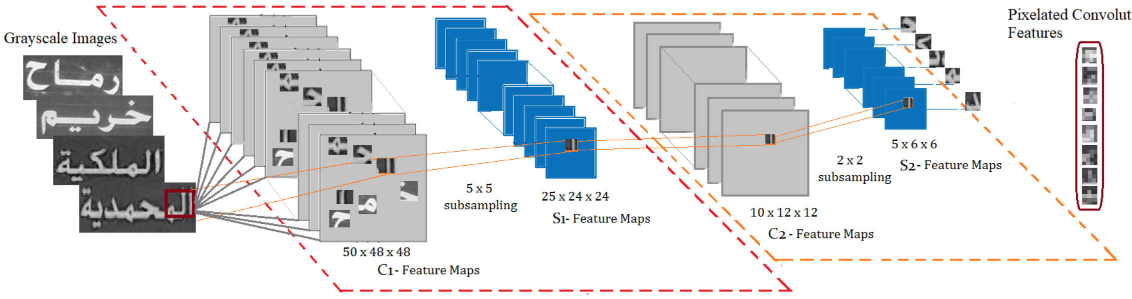


Figure 14. Deep ConvNets' architecture.

The latest work on Latin or English using ConvNets was recently proposed by [61]. They proposed a multi-stage approach for script identification. At first, the text image was normalized by an x-height of 64 pixels while maintaining its aspect ratio, then applying the sliding window approach to extract 32 × 32 image patches, which they called stroke parts. Later, each part was presented to ConvNets to get the feature values. Furthermore, L.G Bigorda et al. [61] explained that each text line is represented as a stroke part's descriptor. They designed the same structure and followed the same pre-processing steps as explained in [62]. They employed the K-means algorithm on extracted patches to learn the convolutional kernels of ConvNets. Instead of extracting convolutional features from a single image with a single vector, as performed by [62], a set of convolutional feature vectors was extracted from parts of the image. Convolutional neural Networks (ConvNets) are another choice to learn the provided patterns, but they are instance learners, which is suitable where data are uncorrelated, like only the characters. If applying ConvNets on extracted image patches, this might not provide stable and good convolutional features because of the inherent complexity attached to cursive scripts. Although this is a good option for feature extraction and mostly applied for feature extraction, they further suggested the idea of merging the Naive Bayes Nearest Neighbour (NBNN) classifier with their proposed solution. The notion behind the said classifier is to compute the image directly without defining intermediate quantized descriptors. Thus, all extracted image patches were given to NBNN to compute the maximum difference of all provided templates that exist in other classes except for its own class. They reported the number of experimental variations on two publicly-available datasets and their own prepared dataset, named the Video Script Identification [8,33] dataset, as well as their proposed dataset for multilingual script named MLe2e, which was comprised of 711 multilingual scene text images. Among various experiments as described in their manuscript, they also performed cross-validation of their trained network on the above-mentioned datasets. The trained network on their proposed dataset MLe2e obtained higher accuracy on the given test images of CVSI and ICDAR 2013, which was 70.22% and 94.70%, while on MLe2e, they obtained 91.67% accuracy.

3.3. Recently-Proposed Deep Learning Research

This section compiles the work to address the issues of scene text analysis presented in 2017 and 2018 using the deep learning architecture.

Several novel deep learning works presented in recent years specifically proposed techniques for correct localization of cursive text. One such work was presented by [79]. The arbitrarily-oriented scene text detection was presented via rotation proposals. By using higher convolutional layers of the network, inclined rectangular proposals are generated with higher accuracy. The pooling strategy was proposed, which is adapted based on rotated Regions of Interest (RoIs). Their proposed technique was evaluated on three real-world text detection datasets, i.e., MSRA-TD500, ICDAR2013 and ICDAR 2015, and obtained good precision, recall and f-measure score. Another novel solution presenting the feature extraction method for scene text extraction was presented by [86]. The super-pixel approach was proposed based on the stroke feature transform approach that was based on deep learning feature classification for text detection. Their proposed technique used deep ConvNets, where each character was predicted by using the pixel value. The hand-crafted features were also used, and they proposed a solution by the fusion of both to obtain high performance system. They evaluated their proposed system on three benchmark datasets, i.e., the ICDAR2011, ICDAR2013 and SVT datasets, and reported the best precision, recall and f-measure score on these datasets. The work on text detection from video images was recently proposed by [80]. They proposed a Bayesian-based network for text detection and recognition. The proposed a system framework composed of three major components, i.e., text tracking, tracking-based text detection and tracking-based text recognition. The details about each category were presented in their article. The detection of text was improved by the proposed multi-frame integration. They evaluated the presented technique on their proposed video text (VidTEXT) dataset collected at University of Science and Technology Beijing (USTB), named USTB-VidTEXT which is publicly available. They reported encouraging results using their proposed techniques. An arbitrarily-oriented text detection by a fully-connected end-to-end convolutional neural network was presented by [81] and also discussed in [65]. The words were predicted by bounding boxes via a presented novel regression model. They evaluated their novel technique on four publicly-available datasets, i.e., the ICDAR 2015, ICDAR 2013, Component Objects in COntext (COCO)-Text images and SVT datasets. They reported state-of-the-art results on publicly-available datasets. The multiple convolutional neural network was proposed by [82]. Their proposed method consisted of three steps, i.e., text-aware, text extraction, text refinement and classification. The proposed architecture by traditional ConvNets, but using multiple layers was used. The proposed technique was evaluated on the ICDAR 2011, ICDAR 2013, ICDAR 2015 and SVT datasets. The details about their performed experiment can be found in their article.

A work on cursive scene text feature extraction was proposed by [83]. The feature extraction approach for Chinese scene text was presented. The features were extracted from the complex structure of Chinese characters by ConvNets. A text structure component detector was presented as one of the layers in ConvNets, which produced robust results on Chinese scene character recognition. The presented technique was evaluated on two Chinese scene text datasets.

By assessing the latest work on scene text analysis as summarized in Table 4, new techniques designed for text detection and classification are presented. Most of the work presented Latin scene text analysis, but few works presented cursive scene text analysis. As this paper is presenting a comprehensive survey on cursive scene text analysis research, therefore, the emphasizes is to investigate the performance of methods presented for scene text detection designed for cursive text, which is very difficult to perform during the whole process of text recognition. In most of the presented techniques, ConvNets were discussed as a base model, which was experimented on by the inclusion of the adapted architecture.

Table 4. Deep learning-based text analysis research presented in recent years. SVT, Street View Text.

Study	Script	Phases	Database
Ma et al. [79] (2018)	Latin	Scene text detection	MSRA-TD500, ICDAR2013, ICDAR2015
Tang et al. [86] (2018)	Latin	Feature extraction	ICDAR2011, ICDAR2013, SVT
Tian et al. [80] (2018)	Latin	Text detection from video images	USTB-VidTEXT
Liao et al. [81] (2018)	Latin	Text detection	ICDAR2015, ICDAR2013, COCO-Text images, and SVT dataset
Tang et al. [82] (2017)	Latin	Feature extraction	7390
Ren et al. [83] (2017)	Chinese	Feature extraction	Own compiled dataset

4. Discussion

This survey summarized the latest trends and techniques that have been applied to cursive scene text recognition, specifically in Arabic script. This section elaborate how the discussed approaches can assist in text detection and recognition of Arabic text appearing in natural images. The prime concern in scene text images is to localize the text in the presence of non-textual patterns. In this survey, the latest trends that were specifically designed for text localization and classification with state-of-the-art techniques are discussed.

The pre-processing of an image is an initial, but essential step, which is considered before applying text localization techniques. The pre-processing requires a homogeneous size and display of an image later, and text localization techniques can be applied to given standardized scene text images. It was learned from the literature that on the basis of scene text complexity, the learning process can further be divided into three major phases (i.e., text localization, feature extraction and text recognition). As noticed in this survey paper, most of the techniques were presented for Latin scene text segmentation, but if one applies these techniques to Arabic script, then it would not produce the expected results. The main reason for this variance in results is the difference of the text's appearance in both scripts.

The reported work on Arabic scene text analysis (as mentioned in the previous section) used manual segmentation by keeping in view the complex structure of Arabic because of the cursive complexity associated with this script in order to obtain high precision in recognition. The subsequent phase after text localization is feature extraction. The feature extraction is meant to observe distinctive properties that exist in the provided sample. In machine learning tasks, the features behave as a backbone for recognition purposes. In this manuscript, the feature extraction approaches defined for cursive scene text recognition are discussed. Being a cursive language and having a complex structure, every character in Arabic has four representations based on its position in a word. At each position, the representation of the character is different. With this constraint, the standardized segmentation approaches may not perform well. In Arabic, the context of character is very important to learn. There is a need to know the previous character so that the next character can be predicted in a word. In this way, sequence learning becomes an integral part of Arabic script analysis. Every acquired text image has built-in features associated with it. These features can be categorized as a unique attribute of an image. As in the feature extraction phase, there is a need to investigate each scene text image so that proper features may be classified during learning. In Arabic, the context is crucial to learn and also to be a part of the given sample during the feature extraction phase. In this paper, some important feature extraction techniques are summarized like global sampling, histogram gradients, scale-invariant transformation and hybrid feature extraction approaches. Among the discussed approaches, SIFT is more relevant to Arabic scene text feature extraction proposed by most researchers with little adaptation. Another very important feature learning approach is window based. By declaring the size of a window in an image, all pixels of an image are considered as a feature regardless of individual character representation. By doing so, the contextual representation of Arabic scene text images is considered.

The extracted features with the corresponding text image are presented to the classifier. The learning of the classification model can be defined as supervised and unsupervised. Most of the work presented in the survey depicts the learning process in an unsupervised manner. Our survey captured the latest trends in cursive scene text recognition by focusing on Arabic scene text. Despite some efforts to make a standard Arabic scene text dataset, we have yet to define a detailed scene text dataset for Arabic script. During classification, such a model should be adopted that can learn the characters with contextual information. In other words, the preference is to learn sequence models for Arabic-like script learning. The best state-of-the-art sequence learner is RNN. The RNN has been adapted to Long Short Term Memory (LSTM) networks, which have been proven to be very useful in Arabic-like text appearing in printed and handwritten script recognition [6,84]. Another very important aspect in relation to Arabic scene text recognition is that the classification techniques based on supervised learning models have not been applied to available datasets.

This survey has provided insight into the on-going research related to Arabic scene text in particular. The details of the collected dataset that is specifically designed for Arabic text analysis in natural images are provided. The overall motive is to provide researchers a viewpoint so that they can assess the potential of their proposed architecture in this unexplored field.

5. Future Directions

The state-of-the-art techniques that have been applied to cursive scene text recognition are summarized in this survey. The emphasis is to have an idea about how research can be performed on recognition of Arabic script appearing in natural images. Future challenges are identified by keeping Arabic script in perspective. Some possible future directions may include:

1. Lack of publicly-available Arabic scene text dataset: Capture and compiling Arabic script scene images is an old challenge itself. The text in natural images usually appears with blur, shudder and at low resolution. Taking an ideal image for research purposes is a provocative task. Although few Arabic scene text datasets are available as reported in [87], to define a detailed dataset in which natural images have text captured in the presence of illumination variation, different text sizes and font styles, there is a need for research on cursive scene text analysis. To create such a dataset, it may prove to be a starting point to address the intrinsic challenges associated with Arabic text in natural images.
2. Localization/detection of text: One of the difficult tasks in scene text images is to localize the text correctly by specific techniques or specialized tools. Extracting text with high precision is still a challenge for researchers to tackle. There exist various approaches to address this issue [4,10,30]. For the purpose of obtaining the segmentation of natural text accurately, Arabic text was localized manually in most of the reported research. A work about text localization of Latin script was presented with high precision by [2], but for cursive script, which is still an open research problem.
3. Text image preprocessing: The scene text image contains unwanted data, which ultimately need to be removed. Most reported work removed such noise manually, or it can be removed by image processing techniques. However, in practical applications when the dataset is large, it is recommended to define methods that may help to make clean text images, as explained in [74]. In that case, an automatic layout analysis tool is preferred, which may detect and remove unwanted information from the given text.
4. Implicit segmentation techniques: Arabic script is very complex in nature due to its various representations of the same characters and ligatures' combination. The association of diacritic marks to a base character is another very important issue. Whilst the automated tools cannot accurately segment this intrinsic script, most of the reported work relied on implicit segmentation of Arabic text. In this regard, LSTM provides an implicit segmentation approach for text recognition as reported in recent research on Urdu character recognition presented by [4–6]. If some research is performed on extensive segmentation approaches, which yields the solution as an implicit method, then it may provide new horizons for researchers who want to exploit their ideas in this direction by using Arabic text recognition in natural images.

5. Apply state-of-the-art deep learning techniques: As mentioned before, most of the proposed Arabic scene text recognition has been exposed to an unsupervised learning algorithm. If applying the state-of-the-art supervised deep learning algorithm for the purpose of estimating correct parameters that are required for training the given sample, then this may lead to the proposal of a new dimension for learning complex patterns like Arabic, as tried by [30]. In this regard, one possible future direction may be the exposure of Arabic scene text images to supervised learning tasks and those deep learning classifiers that perform implicit segmentation of a given text image such as RNNs [7], Convolutional neural Networks (ConvNets) [13] and Bidirectional Long Short-Term Memory networks (BLSTM) [1].

6. Conclusions

This paper aims to summarize recent research that has been performed in cursive scene text recognition specifically in Arabic text in natural images. The most significant and recently-proposed approaches have been explained. Initially, the paper presented details of various cursive scripts; later, the discussion narrowed down with respect to Arabic scene text recognition. Furthermore, this survey examines various text localization algorithms, feature extraction techniques and state-of-the-art classifiers. Most researchers prefer to localize scene text manually in order to get high precision. Another reason for manual extraction of Arabic text is its appearance with various sizes, orientations and font styles. Thus, scale-invariant features are also used because they play an important role in feature extraction. Some authors proposed hybrid feature extraction techniques, which are based on an adapted scale-invariant feature extraction with the combination of their own proposed method. As learned from the presented work, numerous authors proposed unsupervised learning classifiers for Arabic text recognition in natural images. For the purpose of exposing the challenges associated with cursive script, the research under discussion may be exploited by supervised learning methods, so as to obtain a benchmark performance on Arabic scene text recognition research. The discussion of available Arabic scene text datasets including the proposed EASTR dataset was presented. As observed from recently-proposed work, the deep learning architecture is used to address the problems of scene text detection and recognition. Therefore, the research based on deep learning architecture has significant importance in cursive scene text analysis. The future directions were explored by keeping in view the complexities associated with Arabic and Arabic-like scene text recognition after highlighting the current status of cursive scene text detection and recognition research.

Author Contributions: S.B.A. devised a main conceptual idea and worked out almost all of the technical details and contributed in writing of the manuscript with the help of S.N. but under the supervision of M.I.R. and R.Y. The compiled work is verified by M.I.R. who also contributed in some parts of presented contents. All authors provided critical feedback, and helped shape the research, analysis and manuscript.

Funding: This research is funded by Malaysia-Japan International Institute of Technology (MJIIT) under the grant number R.K430000.7743.4J010.

Acknowledgments: The authors would like to thank the Center for Artificial Intelligence and RObotics (CAIRO-ikhoza) lab under MJIIT, Universiti Teknologi Malaysia for arranging funds to conduct this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Graves, A. Teaching Computers to Read and Write: Recent Advances in Cursive Handwriting Recognition and Synthesis with Recurrent Neural Networks. In Proceedings of the CORIA 2014—Conférence en Recherche d'Infomations et Applications—11th French Information Retrieval Conference, CIFED 2014 Colloque International Francophone sur l'Ecrit et le Document, Nancy, France, 19–23 March 2014. Available online: http://dblp.uni-trier.de/db/conf/coria/coria2014.html#Graves14 (accessed on 8 April 2017).
2. Shahab, A.; Shafait, F.; Dengel, A. ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images. In Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 1491–1496.

3. Liwicki, M.; Bunke, H. Feature Selection for HMM and BLSTM Based Handwriting Recognition of Whiteboard Notes. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 907–923. [CrossRef]
4. Ahmed, S.B.; Naz, S.; Razzak, M.I.; Rashid, S.F.; Afzal, M.Z.; Breuel, T.M. Evaluation of cursive and non-cursive scripts using recurrent neural networks. *Neural Comput. Appl.* **2016**, *27*, 603–613. [CrossRef]
5. Naz, S.; Umar, A.I.; Ahmad, R.; Ahmed, S.B.; Shirazi, S.H.; Siddiqi, I.; Razzak, M.I. Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks. *Neurocomputing* **2016**, *177*, 228–241. [CrossRef]
6. Ahmed, S.B.; Naz, S.; Swati, S.; Razzak, M.I.; Umar, A.I.; Khan, A.A. UCOM offline dataset: Aa Urdu handwritten dataset generation. *Int. Arab J. Inf. Technol.* **2014**, *14*, 239–245.
7. Graves, A.; Liwicki, M.; Fernández, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 855–868. [CrossRef] [PubMed]
8. Ul-Hasan, A.; Ahmed, S.B.; Rashid, F.; Shafait, F.; Breuel, T.M. Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1061–1065.
9. Fabrizio, J.; Marcotegui, B.; Cord, M. Text segmentation in natural scenes using Toggle-Mapping. In Proceedings of the 6th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 2373–2376
10. Newell, A.J.; Griffin, L.D. Multiscale Histogram of Oriented Gradient Descriptors for Robust Character Recognition. In Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 1085–1089.
11. Mao, J.; Li, H.; Zhou, W.; Yan, S.; Tian, Q. Scale based region growing for scene text detection. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 1007–1016.
12. Neuhaus, M. Learning Graph Edit Distance. Master's Thesis, University of Bern, Bern, Switzerland, 2003.
13. Ahmed, S.B.; Naz, S.; Razzak, M.I.; Yousaf, R. Deep Learning based Isolated Arabic Scene Character Recognition. In Proceedings of the 1st Workshop on Arabic Script Analysis and Recognition, Nancy, France, 3–5 April 2017.
14. Neumann, L.; Matas, J. Scene Text Localization and Recognition with Oriented Stroke Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 97–104.
15. Simian, D.; Stoica, F. *Evaluation of a Hybrid Method for Constructing Multiple SVM Kernels*; WSEAS Press: Athens, Greece, 2009; pp. 619–623, ISBN 978-960-474-099-4.
16. Tola, E.; Fossati, A.; Strecha, C.; Fua, P. Large occlusion completion using normal maps. In Proceedings of the Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010.
17. Belongie, S.; Malik, J.; Puzicha, J. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 509–522. [CrossRef]
18. Berg, A.C.; Berg, T.L.; Malik, J. Shape Matching and Object Recognition Using Low Distortion Correspondences. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 26–33. [CrossRef]
19. Alginahi, Y.M. A survey on Arabic character segmentation. *Int. J. Doc. Anal. Recognit.* **2013**, *16*, 105–126. [CrossRef]
20. Bissacco, A.; Cummins, M.; Netzer, Y.; Neven, H. PhotoOCR: Reading Text in Uncontrolled Conditions. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 785–792. Available online: http://dblp.uni-trier.de/db/conf/iccv/iccv2013.html#BissaccoCNN13 (accessed on 23 November 2016).
21. Yu, C.; Song, Y.; Zhang, Y. Scene text localization using edge analysis and feature pool. *Neurocomputing* **2016**, *175*, 652–661. [CrossRef]
22. Neumann, L.; Matas, J. Efficient Scene text localization and recognition with local character refinement. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 746–750. [CrossRef]

23. Shekar, B.H. Skeleton Matching based approach for Text Localization in Scene Images. In *Proceedings of the 8th International Conference on Image and Signal Processing*; Elsevier: New York City, NY, USA, 2014; pp. 145–153, ISBN 9789351072522.
24. Liu, X.; Wang, W. An effective graph-cut scene text localization with embedded text segmentation. *Multimed. Tools Appl.* **2015**, *74*, 4891–4906. [CrossRef]
25. Gomez, l.; Nicolaou, A.; Karatzas, D. Improving patch-based scene text script identification with ensembles of conjoined networks. *Pattern Recognition* **2017**, *67*, 85–96. [CrossRef]
26. Raja, S.D.M.; Shanmugam, A. Wavelet Features Based War Scene Classification using Artificial Neural Networks. Scene Classification; Haar and Daubechies Wavelet, 2013. Available online: http://www.enggjournals.com/ijcse/doc/IJCSE10-02-09-104.pdf (accessed on 28 July 2016).
27. Busta, M.; Neumann, L.; Matas, J. FASText: Efficient Unconstrained Scene Text Detector. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2016; pp. 1206–1214, ISBN 978-1-4673-8391-2.
28. Veit, A.; Matera, T.; Neumann, L.; Matas, J.; Belongie, S. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. *arXiv* **2016**, arXiv:1601.07140.
29. Naz, S.; Hayat, K.; Razzak, M.I.; Anwar, M.W.; Madani, S.A.; Khan, S.U. The optical character recognition of Urdu-like cursive scripts. *Pattern Recognit.* **2014**, *47*, 1229–1248. [CrossRef]
30. Parwej, F. An Empirical Evaluation of Off-line Arabic Handwriting And Printed Characters Recognition System. *Int. J. Comput. Sci. Issues* **2012**, *9*, 29–35.
31. Tounsi, M.; Moalla, I.; Alimi, A.M.; Lebouregois, F. Arabic characters recognition in natural scenes using sparse coding for feature representations. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 3–26 August 2015; pp. 1036–1040.
32. Ben Halima, M.; Karray, H.; Alimi, A.M. Arabic Text Recognition in Video Sequences. *Int. J. Comput. Linguist. Res.* **2013**. Available online: https://arxiv.org/abs/1308.3243 (accessed on 20 July 2016).
33. Sharma, N.; Mandal, R.; Sharma, R.; Pal, U.; Blumenstein, M. *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, 23–26 August 2015*; IEEE Computer Society: Washington, DC, USA, 2015; ISBN 978-1-4799-1805-8. Available online: http://dblp.uni-trier.de/db/conf/icdar/icdar2015.html#SharmaMSPB15 (accessed on 14 April 2016).
34. Shivakumara, P.; Sreedhar, R.P.; Phan, T.Q.; Lu, S.; Tan, C.L. Multioriented Video Scene Text Detection Through Bayesian Classification and Boundary Growing. *IEEE Trans. Circuits Syst.* **2012**, *22*, 1227–1235. [CrossRef]
35. Yi, C.; Yang, X.; Tian, Y. Feature Representations for Scene Text Character Recognition: A Comparative Study. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 907–911.
36. Jung, K.; Kim, K.I.; Jain, A.K. Text information extraction in images and video: a survey. *Pattern Recognition.* **2004**, *37*, 977–997. [CrossRef]
37. Pan, Y.F.; Hou, X.; Liu, C.L. Text Localization in Natural Scene Images Based on Conditional Random Field. In Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 6–10.
38. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
39. Yao, C.; Bai, X.; Shi, B.; Liu, W. Strokelets: A Learned Multi-scale Representation for Scene Text Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 4042–4049.
40. Shi, C.; Wang, C.; Xiao, B.; Zhang, Y.; Gao, S.; Zhang, Z. Scene Text Recognition Using Part-Based Tree-Structured Character Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2961–2968.
41. Liu, Q.; Jung, C.; Kim, S.; Moon, Y.; Kim, J.Y. Stroke Filter for Text Localization in Video Images. In Proceedings of the International Conference on Image Processing, Atlanta, GA, USA, 8–11 October 2006; pp. 1473–1476.
42. Tian, S.; Bhattacharya, U.; Lu, S.; Su, B.; Wang, Q.; Wei, X.; Lu, Y.; Tan, C.L. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. *Pattern Recognit.* **2016**, *51*, 125–134. [CrossRef]

43. Pajdla, T.; Urban, M.; Chum, O.; Matas, J. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. *Image Vis. Comput.* **2004**, *22*, 761–767.
44. Chen, H.; Tsai, S.S.; Schroth, G.; Chen, D.M.; Grzeszczuk, R.; Girod, B. Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions. In Proceedings of the 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2609–2612.
45. Gomez, L.; Karatzas, D. Multi-script Text Extraction from Natural Scenes. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 467–471.
46. Yalniz, I.Z.; Gray, D.; Manmatha, R. Adaptive exploration of text regions in natural scene image. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013.
47. Yin, X.C.; Yin, X.; Huang, K.; Hao, H.W. Robust Text Detection in Natural Scene Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 970–983.
48. Zarechensky, M. Text detection in natural scenes with multilingual text. In Proceedings of the Tenth Spring Researcher's Colloquium on Database and Information Systems, Veliky Novgorod, Russia, 30–31 May 2014.
49. Serra, J. Toggle mappings. In *From Pixels to Features*; Elsevier: North Holland, The Netherlands, 1989; pp. 61–72.
50. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006, pp. 404–417.
51. Calonder, M.; Lepetit, V.; Ozuysal, M.; Trzcinski, T.; Strecha, C.; Fua, P. BRIEF: Computing a Local Binary Descriptor Very Fast. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1281–1298. [CrossRef]
52. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision ICCV 2011, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
53. Liu, L.; Wang, L.; Liu, X. In defense of soft-assignment coding. In Proceedings of the IEEE International Conference on Computer Vision ICCV 2011, Barcelona, Spain, 6–13 November 2011; pp. 2486–2493.
54. De Campos, T.E.; Babu, B.R.; Varma, M. Character Recognition in Natural Images. In Proceedings of the Fourth International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, 5–8 February 2009; Volume 2, pp. 273–280.
55. Lucas, S.M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; Young, R. *ICDAR 2003 Robust Reading Competitions*; IEEE: Piscataway, NJ, USA, 2003; pp. 682–687.
56. Olshausen, B.A.; Field, D.J. Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature* **1996**, *381*, 607–609. [CrossRef]
57. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
58. Zhang, Z.; Jin, L.; Ding, K.; Gao, X. Character-SIFT: A Novel Feature for Offline Handwritten Chinese Character Recognition. In Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 763–767.
59. Wu, T.; Ma, S. Feature extraction by hierarchical overlapped elastic meshing for handwritten Chinese character recognition. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK, 6 August 2003; pp. 529–533.
60. Zheng, Q.; Chen, K.; Zhou, Y.; Gu, C.; Guan, H. Text Localization and Recognition in Complex Scenes Using Local Features. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6494, pp. 121–132, ISBN 978-3-642-19317-0.
61. Gomez, L.; Karatzas, D. A fine-grained approach to scene text script identification. In Proceedings of the 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016.
62. Coates, A.; Carpenter, B.; Case, C.; Satheesh, S.; Suresh, B.; Wang, T.; Wu, D.J.; Ng, A.Y. Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning. In Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 440–445.

63. Muja, M.; Lowe, D.G. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In Proceedings of the International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, 5–8 February 2009; pp. 331–340.
64. Wolf, C.; Jolion, J.M. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *Int. J. Doc. Anal. Recognit.* **2006**, *8*, 280–296. [CrossRef]
65. Shi, B.; Bai, X.; Yao, C. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. 2015. Available online: http://dblp.uni-trier.de/db/journals/corr/corr1507.html#ShiBY15 (accessed on 23 December 2016).
66. Boiman, O.; Shechtman, E.; Irani, M. In defense of Nearest-Neighbor based image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
67. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Eventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; pp. 1150–1157.
68. Lazebnik, S.; Schmid, C.; Ponce, J. A sparse texture representation using local affine regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1265–1278. [CrossRef] [PubMed]
69. Varma, M.; Zisserman, A. Classifying Images of Materials: Achieving Viewpoint and Illumination Independence. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2002.
70. Varma, M.; Zisserman, A. Texture classification: are filter banks necessary? In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003; pp. 691–698.
71. Vedaldi, A.; Zisserman, A. Efficient Additive Kernels via Explicit Feature Maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 480–492. [CrossRef] [PubMed]
72. Neumann, L.; Matas, J. A Method for Text Localization and Recognition in Real-World Images. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 770–783, ISBN 978-3-642-19317-0.
73. Eldar, Y.C.; Chan, A.M. An optimal whitening approach to linear multiuser detection. *IEEE Trans. Inf. Theory* **2003**, *49*, 2156–2171. [CrossRef]
74. Hua, X.S.; Wenyin, L.; Zhang, H.J. An automatic performance evaluation protocol for video text detection algorithms. *IEEE Trans. Circuits Syst. Video Tech.* **2004**, *14*, 498–507. [CrossRef]
75. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multi-Oriented Text Detection with Fully Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, CA, USA, 26 July 2016. Available online: https://arxiv.org/abs/1604.04018 (accessed on 21 September 2017).
76. Zayene, O.; Touj, S.M.; Hennebert, J.; Ingold, R.; Amara, N.E.B. Open Datasets and Tools for Arabic Text Detection and Recognition in News Video Frames. *J. Imaging* **2018**, *4*, 32. [CrossRef]
77. Yousfi, S.; Berrani, S.A.; Garcia, C. ALIF: A dataset for Arabic embedded text recognition in TV broadcast. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1221–1225.
78. Slimane, F.; Ingold, R.; Kanoun, S.; Alimi, A.M.; Hennebert, J. A New Arabic Printed Text Image Database and Evaluation Protocols. In Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 946–950.
79. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans. Multimedia* **2018**. [CrossRef]
80. Tian, S.; Yin, X.C.; Su, Y.; Hao, H.W. A Unified Framework for Tracking Based Text Detection and Recognition from Web Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 542–554. [CrossRef]
81. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [CrossRef] [PubMed]
82. Tang, Y.; Wu, X. Scene Text Detection and Segmentation Based on Cascaded Convolution Neural Networks. *IEEE Trans. Image Process.* **2017**, *26*, 1509–1520. [CrossRef] [PubMed]
83. Ren, X.; Zhou, Y.; Huang, Z.; Sun, J.; Yang, X.; Chen, K. A Novel Text Structure Feature Extractor for Chinese Scene Text Detection and Recognition. *IEEE Access* **2017**, *5*, 3193–3204. [CrossRef]
84. Ahmed, S.B.; Naz, S.; Swati, S.; Razzak, M.I. Handwritten Urdu Character Recognition using 1-Dimensional BLSTM Classifier. *Neural Comput. Appl.* **2017**, *30*, 1–9. [CrossRef]

85. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. *arxiv* **2017**, arXiv:1706.09579.
86. Tang, Y.; Wu, X. Scene Text Detection Using Superpixel-Based Stroke Feature Transform and Deep Learning Based Region Classification. *IEEE Trans. Multimedia* **2018**, *20*, 2276–2288. [CrossRef]
87. Tounsi, M.; Moalla, I.; Alimi, A.M. ARASTI: A database for Arabic scene text recognition. In Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, 3–5 April 2017; pp. 140–144.