IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 22, 2018, accepted January 11, 2019, date of publication February 13, 2019, date of current version February 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2895876

A Novel Dataset for English-Arabic Scene Text Recognition (EASTR)-42K and Its Evaluation Using Invariant Feature Extraction on Detected Extremal Regions

SAAD BIN AHMED[1,2], SAEEDA NAZ[3], MUHAMMAD IMRAN RAZZAK[4], AND RUBIYAH BTE YUSOF[1]

[1]Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia
[2]Department of Health Informatics, King Saud bin Abdulaziz University for Health Sciences, Riyadh 11481, Saudi Arabia
[3]GGPGC No.1, Higher Education Department, Abbottabad 22010, Pakistan
[4]University of Technology Sydney, Ultimo, NSW 2007, Australia

Corresponding author: Saad Bin Ahmed (isaadahmed@gmail.com)

This work was supported in part by the Center for Artificial Intelligence and RObotics (CAIRO-ikhoza) Laboratory under MJIIT, Universiti Teknologi Malaysia.

ABSTRACT The recognition of text in natural scene images is a practical yet challenging task due to the large variations in backgrounds, textures, fonts, and illumination. English as a secondary language is extensively used in Gulf countries along with Arabic script. Therefore, this paper introduces English–Arabic scene text recognition 42K scene text image dataset. The dataset includes text images appeared in English and Arabic scripts while maintaining the prime focus on Arabic script. The dataset can be employed for the evaluation of text segmentation and recognition task. To provide an insight to other researchers, experiments have been carried out on the segmentation and classification of Arabic as well as English text and report error rates like 5.99% and 2.48%, respectively. This paper presents a novel technique by using adapted maximally stable extremal region (MSER) technique and extracts scale-invariant features from MSER detected region. To select discriminant and comprehensive features, the size of invariant features is restricted and considered those specific features which exist in the extremal region. The adapted MDLSTM network is presented to tackle the complexities of cursive scene text. The research on Arabic scene text is in its infancy, thus this paper presents benchmark work in the field of text analysis.

INDEX TERMS Cursive script, invariant, extremal, MDLSTM.

I. INTRODUCTION

The field of text analysis in camera captured images constitute a considerable challenge to address by research community. The work presented in recent years, mostly converged on correct detection of text area in presence of other objects in an image [1], [3], [5]. The scene text can be categorized as a typical OCR problem after text detection and segmentation. The presented research on scene text demonstrated good accuracy in a monotonic font but in constraint environment. The scene text usually appears in a varied font and in provocative background that prompt researchers to suggest the solutions to deal with the complexity of cursive script. The text image degradation is another major issue to tackle which is impacted by environmental constraints in natural images such as, text orientation, illumination and blurry effect. The features always determine distinct characteristics associated within an image. The determination of relevant features from text area is a potential problem to inscribe. In addition to varied text sizes appeared in scene images, the other challenging tasks scene text may suffer are text orientation, color and layout complexities. These tasks may influence and degrade the text appearance by illumination factor as depicted in Figure 1. In this figure, the text contour detection by Sobel filter represents Arabic text appeared in various orientations, blur and in different font style. Sometimes the noise is represented at a same energy level which can lead to miss-classification of a text as shown in Figure 2. The processed images presented in the figure are generated by applying Sobel filter. The red box in an image as represented in the

The associate editor coordinating the review of this manuscript and approving it for publication was Imran Ahmed.

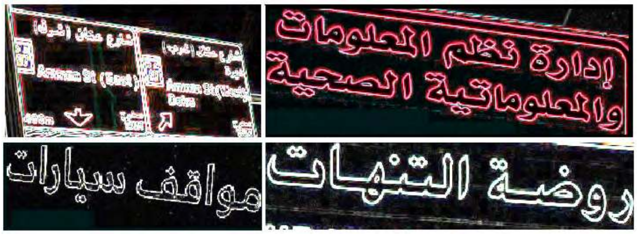


FIGURE 1. Arabic scene text variations.

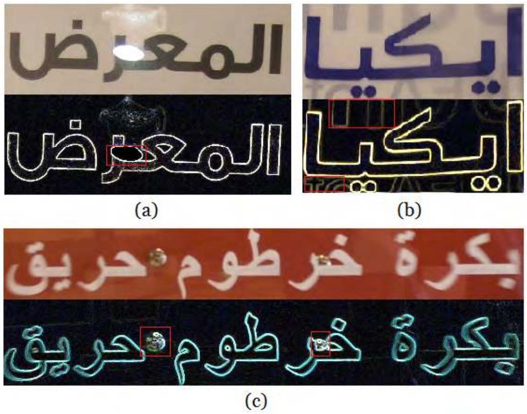


FIGURE 2. Challenging scene text images.

figure indicates the noise that can be detected as a part of a text.

The contemporary feature extraction approaches proved to be very accurate and focused on text area regardless of implicit ambiguity and distortions associated with these kind of images as represented in [6] and [9]–[11]. The text localization is considered computationally very difficult and expensive task in terms of efforts. Generally, it requires 2^N subset for a single text (N is representing number of pixels over text area). To find the individual characters by merging those pixels that share similar attribute together and make a region by using a technique such as connected components analysis, is commonly experimented to find text in camera captured images. The most prominent method that has been experimented by number of researchers is Maximally Stable Extremal Regions (MSER) as represented in [7], [13], and [14]. The significant results are reported by considering this technique on scene text localization but there is a need to apply post processing strategies so that non-text regions can be eliminated.

A. COMPLEXITIES RELEVANT TO SCENE TEXT IMAGES

Usually in scene text images, the influencing factors that make effect over the recognition accuracy are illumination, varied font size and styles, text texture, color blending, and text orientation create problems during text classification. The aforementioned complexities are not linguistic specific rather to deal with such challenges make it convenient for correct localization of other cursive text. Therefore in literature, by considering the complexities of relevant script, various techniques have been proposed separately for text localization, extraction and recognition as explained in [6]–[9].

Many ways are defined to deal with the complexity of scene text especially in text localization. One such approach is region based method proposed by Pan *et al.* [12] which relies on the assumption that text regions in images have clear-cut characteristics in comparison to non-text regions. The text edges are examined and then can be extracted from the natural image. During the text detection phase, every region is evaluated on the assumption of having a text. Text localization techniques like connected component analysis, color dissimilarity information, or contour detection applies on scene text image for text extraction purpose. Still, to acquire accurate position of a text is a challenge for researchers as the text recognition accuracy is not better in determination of text region in an uncontrolled environment. The efficiency is another issue pertaining to connected component analysis as it consume reasonable time to figure out correct text segmentation due to orientation sensitivity and text visibility caused by illumination factors or other impediments.

Since few years, it is observed from presented literature that region based technique i.e., MSER, is widely experimenting with combination of other techniques as proposed by Elnemr [7] Bigorda and Karatzas [13]. The MSER assumed that text color has high contrast in comparison to it's background in an image. The text localization is performed by MSER to search for homogeneous color region in an image. In MSER, the hue (H), saturation (S), and Intensity (I) are accounted with combination of gradient channels for the purpose to detect regions or components. In this paper, the proposed work is extracting invariant features from detected extremal regions. The scale invariant feature keypoints are spotted on extracted text regions. The extracted keypoints regarded as highly degenerated points but have low discrepancy with respect to scale. The proposed work exploits the advantage of scale invariant approach and applied it on extremal text regions. The extracted features of detected text region have distinctive patterns that is trained by proposed Long Short Term Memory (LSTM) neural network based classifier. The detail about LSTM is mentioned by Graves [15].

This paper proposed novel technique by adapting the scale invariant approach and applied it on detected text regions in natural images. The invariant features are extracted because the extracted keypoints from reference images are invariant to transformation factor. In Arabic, there is complex and complicated text appeared in uncontrolled environment which needs to be addressed by using the robust techniques such as scale invariance method. Another aspect that is covered by proposed methodology is to focus on area of interest i.e., text region. The proposed work suggested, that it is more appropriate to use MSER first on reference images then apply invariant feature technique for a purpose to extract keypoints. After extraction of keypoints, the concentration is given on those relevant keypoints that appeared in conjoined area with reference to binary image and image mask.

The Arabic scene text dataset is also proposed by Tounsi *et al.* [16], but their dataset is limited in number and also lack detailed variations of images in comparison to proposed dataset which covers huge number of Arabic scene text samples. The images were captured from different sources as indicated in section III. The proposed technique has applied on Arabic Scene text dataset which is evaluated by multidimensional LSTM based classifier with the accuracy nearly 94.01% which is reported by considering selected keypoints. The hybrid approach based on scale invariant feature extraction or local region extraction by MSER has not been addressed before. By detecting the region, the proposed work is precising the effort in determining the localized text region. In this way, varied amount of features are extracted in relevance to provided text image. In addition to Arabic script, the evaluation is also performed on captured English samples. The features from English samples are extracted by following window based approach with combination of MDLSTM architecture.

The proposed method for Arabic scene text recognition is not any language specific rather has a potential to be applied on any cursive scene text image appeared with implicit impediments. The method adopted for Arabic scene text recognition can be applied on any cursive script. In this case the procedure will be the same for recognition process but dataset will be different. For English samples, the proposed technique can be evaluated on other English/ Latin scene text dataset.

The contribution in this paper is to propose a recognition system described as follows,

1) Invariant features from each detected extremal regions are determined in binary image and in image mask. The co-occurance of invariant features in both images are considered as potential features.
2) The adapted LSTM network is trained on extracted potential concurrent features.
3) The LSTM network is evaluated on proposed English-Arabic Scene Text Recognition (EASTR) dataset.

This paper focuses on proposing hybrid approach for Arabic and English scene text detection and feature extraction. The MSER is suitable for detecting text area from a given image but most of the time it miss-classifies the text, so there is a need to manually verify the text images as extracted by MSER approach. In addition, the details about proposed EASTR-42K dataset is also presented. The recurrent neural network approach on proposed dataset is evaluated to assess the potential of presented dataset.

II. RELATED WORK

The various efforts have been reported in the field of scene text recognition during last decade, most of them highlighted the issue of text localization and presented solutions for it. The word *text localization* and *text detection* use interchangeably in this manuscript.

The numerous methods are proposed to extract the text from camera captured images. The acquired images can be categorized as an image of license number plate, an image of road guide text, text appeared on advertisement board, image of a pamphlet, and a text written on various commodity wrappers. Yet, there is a need to propose generalized method defined for text extraction presented with implicit constraints. The camera captured text images are not taking in a controlled environment therefore, heterogeneity of challenges are associated with correct detection of a text. This section is presenting the compiled work which is based on recently proposed approaches for scene text localization specifically designed using MSERs and scale invariant feature extraction technique.

A generalized approach for text localization and recognition is proposed by Neumann and Matas [17]. They adapted MSER for text detection/ localization which provide geometric precision to locate text region in an image. Furthermore, they proposed hypothesis verification framework and train the algorithm on synthetic fonts. They evaluated the performance of proposed technique on two benchmark datasets i.e., Char74k and ICDAR2003 dataset. Although they reported good accuracy on Char74 i.e., 72% recognition rate, but on ICDAR2003 57% recognition rate was reported which considered as a worse recognition rate in comparison to other techniques but as they mentioned that they evaluated whole ICDAR2003 dataset on their proposed technique which has not been experimented before. Another very interesting work is represented by Fabrizio *et al.* [18], they proposed a toggle mapping technique to segregate the text from natural scene images. Initially, this method was used for contrast enhancement and noise reduction. The grayscale image is segmented into two functions, further they performed morphological erosion and dilation on each function. The overall goal is to detect the boundary of a text. The homogeneous region in an image may be the point of interest. They considered two functions of grayscale and a minimum contrast value for the purpose to locate text in an image. They evaluated their proposed method on 501 readable characters and found 74.85% accuracy on correctly segmented characters. A hybrid method for scene text detection in real time environment is presented by Bigorda and Karatzas [13]. They termed their proposed technique as a performance optimized. Their proposed module is based on MSER which detect and track the text asynchronously. The technique works on two steps, the text detection by MSER is performed as a first step while in second step they find region of interest in successive frames where there exist a correspondence problem. Moreover, the mismatch is detected by RANSAC algorithm that is explained in [19]. The OpenCV Android framework development is used for their proposed method implementation. Their proposed system is evaluated on variable frame rate, the average frame rate is 25 frame per second (fps). They find that time performance in tracking a module would increase linearly with respect to detected regions and size of their search engine.

A work on arbitrary orientation of an image is proposed by Yao *et al.* [20], they suggested scale invariant based technique on extracted component and perform analysis on them. Later, they link the candidate region and verify the process by chain analysis. On the basis of analysis, they designed two set of features i.e., component level and chain level features. After determining the irregularities in an image, text and non-text area was disintegrated. In order to determine the text, they observed the constant width, texture lessees and smoothness of strokes. They also proposed a multilingual dataset which is divided into train set and test set. The name of their dataset is MSRA-TD500. Their dataset contains 500 images among them 300 used for training purpose and remaining 200 used to test the performance of proposed method. The performance of various text detection methods was investigated on ICDAR dataset and also yielded the performance by evaluating different text detection algorithms on their proposed dataset which produced good results in comparison.

There are several novel work presented in recent years that specifically proposed techniques for correct localization of cursive text. One such work is presented by Ma *et al.* [21]. They presented arbitrary oriented scene text detection via rotation proposals. By using a higher convolutional layers of network inclined rectangular proposals are generated with higher accuracy. They also proposed a pooling strategy which is adapted according to rotated region of interests RoIs. They evaluated their technique 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 feature extraction method for scene text extraction is presented by Tang and Wu [22]. They proposed super-pixel based stroke feature transform approach that is based on deep learning feature classification for text detection. Their proposed technique is based on deep ConvNets where each character is predicted by using the pixel value. They also used hand-crafted features and proposed solution by fusion of both to get high performance system. They evaluated their proposed system on three benchmark datasets i.e., ICDAR2011, ICDAR2013 and SVT datasets and reported best precision, recall and f-measure score on these datasets. The work on text detection from video images is recently proposed by Tian *et al.* [23]. They proposed Bayseian based network for text detection and recognition. The proposed system framework proposed three major components i.e., text tracking, tracking based text detection, and tracking based text recognition. The details about each category is mentioned in their article. The text detection is improved by proposed multi-frame integration. The evaluation is performed on their proposed dataset named USTB-VidTEXT which is publicly available. They reported encouraging results using their proposed techniques. An arbitrary oriented text detection by fully connected end-to-end convolutional neural network is presented by Liao *et al.* [24]. The words are predicted by bounding boxes via a presented novel regression model. They evaluated their novel technique on four publicly available datasets i.e., ICDAR 2015, ICDAR 2013, COCO-Text images, and SVT dataset. They reported state-of-the-art results on publicly available datasets. The multiple convolutional neural network is proposed by Tang and Xiangqian [25]. Their proposed method consists of three steps i.e., text-aware, text extraction, text refinement, and classification. They proposed architecture by traditional ConvNets but using multiple layers. The proposed technique is evaluated on ICDAR 2011, ICDAR 2013, ICDAR 2015 and SVT Datasets. The detail about their performed experiment can be found in their manuscript. The cursive scene text feature extraction is presented by Ren *et al.* [26]. The feature extraction approach for Chinese scene text is presented. The features from complex structure of Chinese characters were extracted by ConvNets. A text structure component detector is presented as one of the layer 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 presented for scene text analysis, it is noticed that new techniques designed for text detection and classification are presented. Most of the work presented for Latin scene text, but few work presented for cursive scene text analysis. As this paper is presenting 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, the ConvNets is discussed as a base model which is experimented by inclusion of adapted architecture. This paper presented novel feature extraction approach designed specifically for the Arabic text in natural images.

The rest of the paper is organized in different sections. In section III, the characteristics of collected data samples, pre-processing steps, generation and validation of ground truth are presented. Section IV highlights the importance of context learning in perspective of cursive script. The presented methodology is explained in section V. Section VI exhibits the evaluation results whereas, comparison of proposed method performance is detailed in section VII. The summary of presented work is depicted in conclusion section.

III. ARABIC SCENE TEXT ACQUISITION AND STATISTICS

This section provides detail about collected data samples having Arabic text. The characteristics of Arabic script, process of acquiring data samples, preprocessing of captured text samples and generation and verification of ground truth are described in this section.

A. ARABIC SCRIPT PROPERTIES

Arabic is considered as an ancient script which is followed by many languages i.e., Arabic, Persian, Urdu etc. It is rich in vocabulary and is used by 1/3rd of world population.

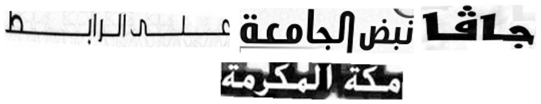


FIGURE 3. Variations in writing style.

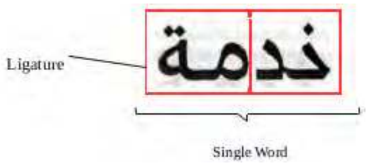


FIGURE 4. Ligature representation in Arabic word.

Arabic script is purely cursive which exhibits various forms of single character depending on the position it occurs in a word. Cursive, context sensitive and calligraphic nature of this script make it quite challenging to classify especially in scene text images as shown in Figure 3. Shape of the character depends on preceding and following character in the ligature (i.e., initial, middle, final or isolated). The ligature plays an important role in word formulation, moreover one word may consist of one or many ligatures. Unlike, Latin and other scripts, Arabic script is written from right to left. Figure 4 shows two ligatures which make a single word pronounced as 'khidmath'. The red box represents the ligatures which is a part of a word. Another property of Arabic text is representation of joiner and non-joiner characters. The joiner character occurs at final, isolated, initial position or at middle position and it may completely change its shape at middle position and initial position. Whenever, non-joiner character appears at final position in a word or as isolated form, it must terminate the word. The end character of a word maintains its full shape.

B. DATASET COLLECTION

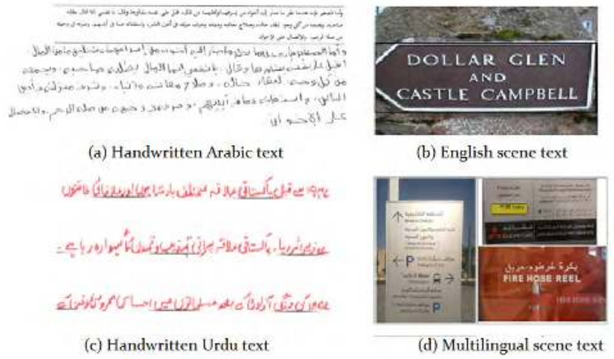
The multilingual scene text images are captured by specialized cameras . The multilingual nature of acquired samples prompted to compile a dataset for English in parallel. The database collected for scene text is divided into three sections i.e., Arabic, English and for multilingual scripts. The acquired scene text images were taken from University precinct, advertisement boards, guide boards displayed on roadsides and also taken from various commodity wrappers. The 2469 scene text images have been captured which comprise number of text lines and Arabic numeral. The Arabic scene text data is categorized into text lines, words and characters. All images were taken from *NikonD*3300 specialized DSLR camera with 24 mega pixels lens and from HTC-One(M8) with 2.5 GHz quadcore along 2GB RAM which has 13 mega pixels back camera with ultra pixel sensor, it means that it may capture more light. All images were captured in an uncontrolled environment. The image dimension is 2688 × 1520 with fixed 72 dpi at horizontal and vertical resolution. The exposure time is varied according to the light expose on the object.

TABLE 1. Sources of acquired images.

Source	Number of acquired test images
University precinct	300
Hoardings	383
Roadside guide	1539
Commodity wrappers	247

C. MULTILINGUAL SCENE TEXT RECOGNITION AND ITS NEED

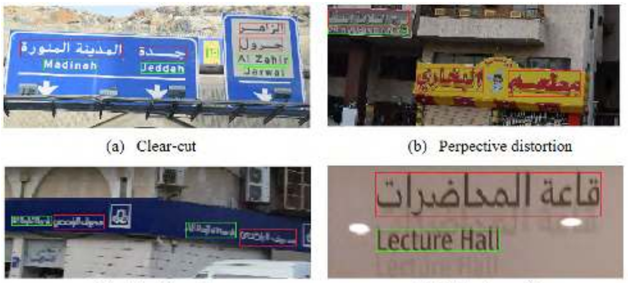
Recent development for scene text recognition mostly focused on Latin or English text. However, few efforts have been done for cursive script scene text recognition specifically the Arabic. Even though some efforts have been made on different script, but some raised question need to be answered. Among them the foremost question is relevant to the need of multilingual scene text recognition system. Generally, character recognition systems are uni-language as most of the text exist in single language as shown in Figure 6. However, it is not true with scene text images. In non-English countries such as Arabian peninsula and Indian subcontinent, mainly scene text images appeared in multilingual scripts as shown in Figure 5. With emerging multilingualism techniques, bilingual, trilingual or even more languages need to be supported. English is one of the main language used with other local script i.e., Arabic and Urdu. Thus, there is an urge



(a) Handwritten Arabic text (b) English scene text

(c) Handwritten Urdu text (d) Multilingual scene text

FIGURE 5. Cursive and non-cursive handwritten/ multilingual scene text and data samples. (a) Handwritten Arabic text. (b) English scene text. (c) Handwritten Urdu text. (d) Multilingual scene text.



(a) Clear-cut (b) Perpective distortion

(c) Jerky-blurred (d) Reflective surface

FIGURE 6. Sample images in EASTR dataset. (a) Clear-cut. (b) Perpective distortion. (c) Jerky-blurred. (d) Reflective surface.

to develop multilingual scene text recognition system which can work seamlessly across different scripts.

To propose a dataset having the Arabic script in focus is a worthily effort in terms of its size that mainly covers more variations with respect to orientation, illumination and font styles. Many scene text datasets are discussed as reported in ICDAR competitions [27]–[30], but the scene text dataset for the Arabic requires comprehensive representation of the Arabic text. After surveying available literature, it is learned that only one camera captured Arabic scene text dataset named ARASTI proposed by Tounsi *et al.* [16]. They presented limited number of acquired samples having less variation in taken sample. Moreover, the dataset covered images taken from commodity wrappers. The images of pamphlets and books presenting Arabic text and numerals also part of proposed dataset. Arabic text itself considered as complex due to its cursive nature which accentuate numerous challenges during recognition. The ASTR dataset contains every possible word appears in Arabic language with all it's permutation in reference to shapes, fonts, size of a text and font colors. In general, ASTR dataset covers huge variety of Arabic scene text appeared in unconstraint environment. The acquired dataset named "EASTR" as it covers English words in addition to Arabic. The details about EASTR dataset is briefly elaborated below.

D. EASTR-42K DATASET

This section details about the EASTR-42K dataset collection process and its statistics. The proposed data captured bilingual (English and Arabic) scene text images and tried to cover every possible word permutation of Arabic language with it's variant shapes. The acquired text images were segmented into English and Arabic text lines, words and characters. The segmented words from Arabic text lines are depicted in Figure 7, while segmented Arabic characters represented in Figure 8. Due to different font styles and cursive nature of Arabic script, there is a need to have such dataset which covers maximum number of Arabic text so that it may consider all possible Arabic variations. The EASTR-42K dataset covers huge variety of English and Arabic scene text appeared in unconstrained environment. The details about EASTR-42k dataset division based on complexity, total number of images, text lines and segmented words and characters are briefly elaborated in Table 2,3,and 4. In Table 2, the detail about division of collected samples based on complexity with respect to Arabic and English is mentioned. The acquired complex text is divided into text lines, words and characters.

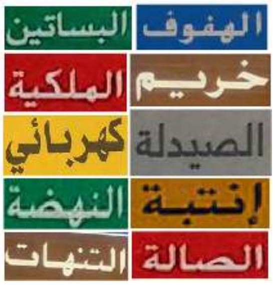


FIGURE 7. Sample word images in EASTR dataset.

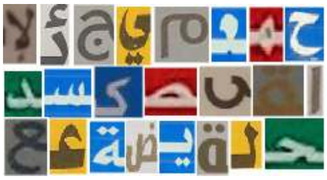


FIGURE 8. Segmented characters representation in EASTR dataset.

TABLE 2. EASTR-42K division based on complexity.

EASTR-42K Dataset			
Language	textlines	words	characters
Arabic	8915	2593	12000
English	2601	5172	7390

TABLE 3. Number of text lines in Arabic, English and Multilingual.

Language	Number of Textlines
Arabic	2107
English	983
Multilingual	784

TABLE 4. Number of characters assuming 6 character per word in Arabic and English.

Language	Number of Characters
Arabic	16624
English	5904

Table 3 depicts the total number of good samples exist in EASTR-42k dataset including multilingual text image covering Arabic and English. Table 4 summarizes the description about number of Arabic and English characters appeared in better quality acquired scene text images.

E. PREPROCESSING OF SCENE TEXT IMAGES

In Arabic, it is cumbersome to disintegrate word into individual characters as discussed earlier. Furthermore, in cursive style, it is impossible to correctly segment the characters. The character shape variation, its position and occurrence of two consecutive characters on same level, makes a challenge for segmentation techniques to work perfectly on complex scripts like Arabic. In this scenario, implicit segmentation plays its role that segments the characters empirically. The quality of text presented in captured images were impacted by the presence of illumination factor appeared in uncontrolled environment. Such illumination factor may cause implicit noise attached to acquired samples which ultimately may blur the visibility of a text.

The unnecessary data should remove prior to classify them. The scene text image is manually segmented into different text lines for example, into 6 text lines as represented in Figure 9. The skew is detected and corrected empirically

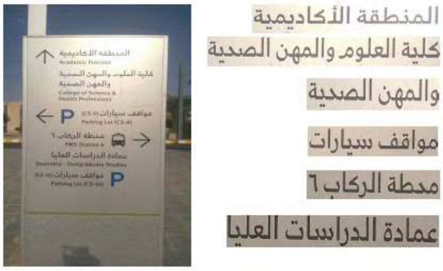


FIGURE 9. Captured image with segmented Arabic textlines.

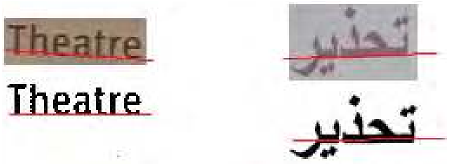


FIGURE 10. Image representation with skew correction.

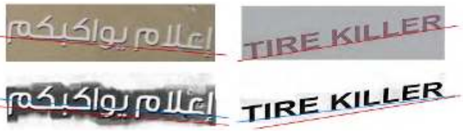


FIGURE 11. Large skewed image.

by RAST based skew detection and correction technique using OCRopus as explained in [31] and shown in Figure 10. The skew correction is done empirically but there is largely skew images present in EASTR dataset. The recognition results are deviated on largely skewed text images as observed in Figure 11. The image is represented with large skew. The red line indicate actual skew, whereas the blue line shows corrected skew. The acquired images having Arabic text are standardized by considering the x-height. The MSER technique is applied afterwards, to determine the region of interest. Subsequently, SIFT method is used to consider invariant features in extracted region by MSER. The aspect ratio of an each image is maintained by keeping in view the varied width of a text image. The subsequent step is generation of ground truth labels as explained in following sub-section.

F. GENERATION AND VERIFICATION OF GROUND TRUTH

The establishment of ground truth declaration is crucial step in supervised learning methods. It is considered as one of the salient step to match the learned pattern with target value. The determination of ground truth solely depends on relevant classification technique and implementation scenario. The ground truth for Arabic script is divided in two different ways. This manuscript is presenting hybrid feature extraction approach by combination of the extremal regions and invariant features. The extracted regions are labeled and consider only those x-y coordinates which mutually occurred in both binary image and an image mask.

Another way to establish a ground truth is label based, it is used to define each word in a dataset. The ground truth is written in text file accompanying with its image file with same name. There are 27 basic characters and 10 numeral in Arabic language. For ground truth generation, 37 classes are declared, every class is a representation of every single character regardless of its position occurred in a word. The Latin characters used to write a ground truth file. The ground truth file is manually generated as depicted in Figure 12. As shown in this figure, there are 8 characters used in making two words. The ground truth is labeled as represented in Figure 13, every character was separated with '–' symbol and word separated by space in-between two words. Each class is defined using Latin characters. The ground truth file is manually verified for any correction. The missing character verification is automatically detected by written program. The missing characters are those which may be specified in ground truth but not declared as a class during implementation. The context is treated as an integral part in cursive text recognition. The representation of each character depends on its previous character and so on. In next section, the importance of implicit segmentation and context learning of cursive script is contemplated.

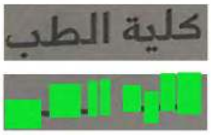
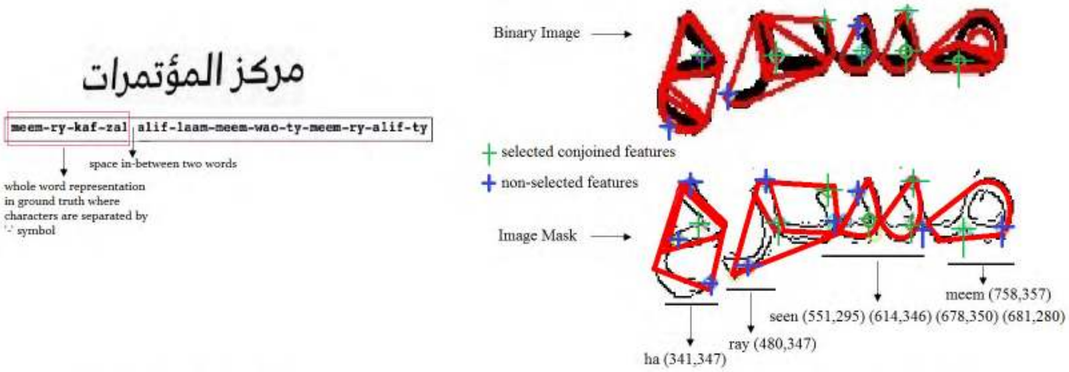


FIGURE 12. Ground truth depiction.

IV. CONTEXT LEARNING WITH PERSPECTIVE OF IMPLICIT SEGMENTATION

In cursive scripts like Arabic, an implicit segmentation provides a way to segment the characters correctly by using techniques like dynamic programming. In Arabic, there are variations of same characters with respect to its position in a word. In such complex script, it seems cumbersome to recognize a same character having variant shapes. In these intrinsic scripts, context plays a major role in recognition. The context learning approach named Multidimensional Long Short Term Memory (MDLSTM) network is discussed by Graves [15]. The MDLSTM is considered as connectionist approach which mainly relies on Multidimensional Recurrent Neural Network (MDRNN) and Long Short Term Memory (LSTM) networks. The MDLSTM follows RNN approach to learn the sequences. All past sequences with respect to current point in time are accumulated to predict the output character.

The RNN is most suitable for sequence learning tasks. The temporal sequential behavior of RNN is recorded by Connectionist Temporal Classification (CTC). By nature, CTC is suitable to use with bidirectional model for the purpose of making estimation on both sides with reference to current point during processing. As explained in [6], general neural network with an objective function requires separate training targets for every segment or time step with the input sequence. This leads to two major implications.


مركز المؤتمرات
meem-ry-kaf-zal
alif-laam-meem-wao-ty-meem-ry-alif-ty
space in-between two words
whole word representation in ground truth where characters are separated by '-' symbol
Binary Image
+ selected conjoined features
+ non-selected features
Image Mask
meem (758,357)
seen (551,295) (614,346) (678,350) (681,280)
ray (480,347)
ha (341,347)

(a) Groundtruth generation from Latin characters

(b) Groundtruth generation by conjoined invariant features detection in extremal regions

FIGURE 13. Two different ways of ground truth declaration. (a) Groundtruth generation from Latin characters. (b) Groundtruth generation by conjoined invariant features detection in external regions.

1) The data samples for training should be pre-segmented.
2) The local classification is computed against every input label, but global aspect is missing which is required to make the context in case of large and complex sequential problems.

To address above mentioned issues, it is revealed that the trained network requires sequences to predict the final character or symbol with perspective of global aspect. The CTC technique addresses sequence problem. The alignment of a sequence is no more important after inclusion of CTC in RNN architecture, because it can make prediction at any point in time against input labels until the whole label's sequence is correct. The CTC is added as an output layer in RNNLIB library as a softmax layer. To complete the sequence, probability is estimated that consequently eliminate the efforts require for post-processing as highlighted by Halima *et al.* [32]. The further detail about CTC, its working, and importance with respect to recurrent neural networks can be found in [15].

V. METHODOLOGY

An adapted MSER text detection method is proposed. This method extracts invariant features from detected text blob. The presented classifier learns the specific extracted invariant features which are instrumental during recognition process. The idea is presented in following subsections.

A. DETECTION OF EXTREMAL REGIONS

The well known connected component approach for detection of extremal region is maximally stable extremal region approach. The word extremal points towards the characteristics where all pixels inside MSER detected blob would have higher or lower intensities in comparison to its outer boundary. The MSER detects co-variant points and merge them together to make a region. The main idea of MSER is to detect those points in an image that stay nearly the same in presence of wide range of thresholds. The regions having minimum variations at the time when threshold applied would be considered as maximally stable regions. During the process, over the large threshold the binarization of image is stable which means that it represents minimum invariance to affine transformation with respect to involved pixel intensities as represented in following equation.

$$f(x) = Ax + T \tag{1}$$

$f(x)$ is affine function which represents linear attitude A and a transformation variable T.

The whole image is evaluated by following equation,

$$f(x_i) = \sum_{i=1}^{n} Ax_i + \sum_{i=1}^{n} T_i \tag{2}$$

The extraction steps of MSER are mentioned as follows,

1) Apply threshold algorithm over the whole image.
2) Find extremal regions by connected component analysis.
3) By threshold, the maximally stable regions is detected in an image having discrete nature.

It is pertinent to mention here that extremal region might be rejected. The rejection is based on the fact, if detected region covers maximum area or minimum area. Furthermore, if region is unstable and a possibility of having duplicate extracted regions, then rejection will take place. The important characteristic of extremal region is the continuous affine transformation, hence this feature some times could not be able to extract exact region of interest that requires considerations from research communities to work on for further precision. In presented work, stable regions needs to be searched out by evaluating binary image and an image mask as shown in Figure 15. The extraction of interested regions depend on image quality. As observed in Figure 15, the precision of text detection is clearly visible in binary image as compared to image mask. But there are some situations where text is localized precisely in an image mask in comparison to binary image, hence the quality of an image plays a vital role for text localization. Therefore, the proposed work validates on most of the presented images. Moreover, it concluded with

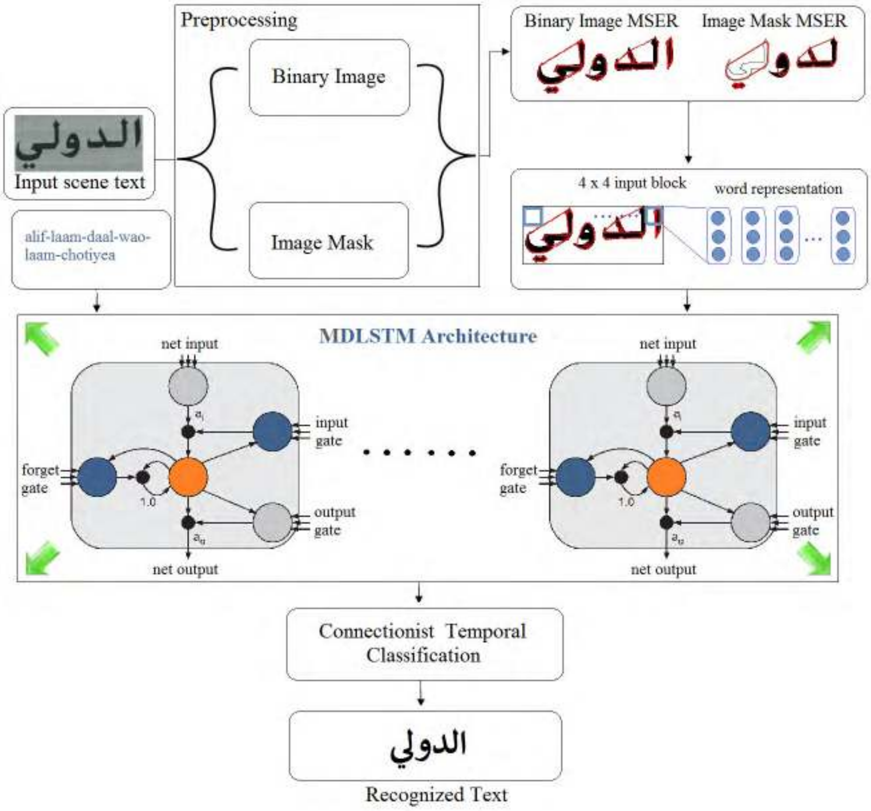


FIGURE 14. Proposed system architecture.

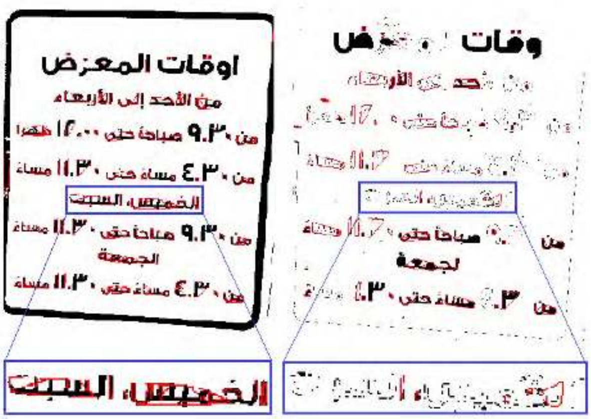


FIGURE 15. (a) Extremal Regions based Text detection in binary images (on left). (b) Text detection in image mask (on right).

the established fact that binary image enhances the quality of an input image which eliminates unnecessary details that may hinder the performance otherwise.

The non-interested regions can be minimized through transformation by applying image filtration methods. The intent is to detect text in a scene image, but there are some non-text regions which should be treated as noise. The overall text detection yielded very good accuracy for Arabic script in particular.

Algorithm 1 Algorithm for Text Localization

```
Input: A filtered Image
Output: Text detected Image
    Procedure for Text Localization :
 1: Take a raw Image $I$
 2: Perform filtration technique on Image $I_B$
 3: where, $I_B(x,y) = \begin{cases} 1, & \text{for } I_B(x,y) > t \\ 0, & \text{for } I_B(x,y) \leq t \end{cases}$
 4: Apply adapted-MSER approach on $I_B$.
 5: An adjacency relation is defined on $I_B$
    IR is Maximally Stable Region
 6: Let $IR_1, ...., IR_{i-1}$ is a sequence of nested extremal
    regions $\mathbb{R}$
    LOOP Process
 7: for $i$ = no. of $IR$ do
 8:     if i == $\mathbb{R}$ has a local minimum at i* then
 9:         Select $\mathbb{R}$ as a candidate region
10:         if i $\neq$ 0 then
11:             Goto step 8, unless $i$ does not possess any value
12:         endif
13: end for
```

As mentioned in an Algorithm 1, an adjacency relation is defined on images.

Suppose, two different areas are $p, q \in D$, here D belongs to entire image, wheres IR is a subset of D. p is adjacent to q $p \forall q$, if,

$$\sum_{i=1}^{d} |p_i - q_i| \tag{3}$$

where i is a number of regions in an image. For each $\mathbb{R}$, there is a sequence p like, $a_1, a_2, \ldots..a_n$.

B. INVARIANT FEATURE EXTRACTION

The scale invariance is an eminent feature of scale invariant feature transformation (SIFT) method. To achieve scale invariance, SIFT uses Laplacian pyramid which is calculated by difference of various level of Gaussian (DoG) function as represented in equation 4.

$$D(x, y, \delta) = (G(x, y, \delta_k) - G(x, y, \delta)) * I(x, y) \tag{4}$$

where,

$$G(x, y, \delta) = \frac{1}{2\pi e\delta^2} exp[-\frac{x^2 + y^2}{2\delta^2}] \tag{5}$$

By Laplacian pyramid L as represented in equation 6, high frequency information of an image can easily be obtained, because features in an image mostly resides on these parts.

$$L(x, y, \delta) = G(x, y, \delta) * I(x, y) \tag{6}$$

The scale space is divided into various octaves. In each octave the initial image is convolved with the Gaussian G to produce the set of scale space. The adjacent Gaussian is subtracted to get difference of Gaussian. After each octave, the Gaussian image is down sampled by factor 2 and rest of the process is repeated in the same manner. The number of octaves helps in finding the key points in different scales. The octave number and scale depends on the size of an original image.

During determination of keypoints, at each level of DoG octave, the pixel point I(x,y) is investigated whose value should be greater(or smaller) than eight adjacent pixels. At these extracted points, the value of adjacent pixels resides in lower and upper level are compared. In first and last scale, there are not enough adjacent pixels to compute which restrict in finding local minima or maxima. Based on the criteria, minimum or maximum value, location, and scale of a relevant point, are recorded. The detection of all keypoints in an image do not imply that it will contribute in recognition, instead there is a need to accept or reject unnecessary feature points which are generated on low contrast region and are poorly localized along the edge. Therefore, it is assumed that all extremal points extracted through DoG space search help in finding location, scale and orientation of each keypoint. To obtain consistent orientation with respect to each extracted keypoint based on local image properties, a keypoint descriptor is defined to represent orientation information.

Suppose, the orientation of detected keypoints is assigned as shown in Figure 16 and mathematically expressed in equation 8. The region is selected having a keypoint in the center. The region size is covered within the circle where keypoint is detected as shown in Figure 17 and mathematically represented through equation 7 and 8.

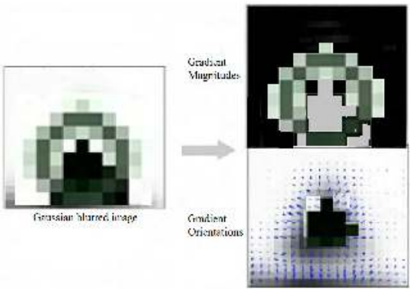


FIGURE 16. Gaussian blurred image represented as Gradient magnitudes and orientation.

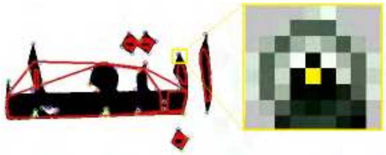


FIGURE 17. Detected keypoint.

$$\begin{aligned} B(x, y) \\ = \sqrt{(L(x+1, y) - L(x-1, y)^2) + (L(x, y+1) - L(x, y-1))^2)} \end{aligned} \tag{7}$$

$$\begin{aligned} \theta(x, y) \\ = tan^{-1}((L(x, y+1) - L(x, y-1)) \\ /L(x+1, y) - L(x-1, y))) \end{aligned} \tag{8}$$

The next step is to define the image descriptor which contains all information of extracted keypoints that are categorized as a distinctive feature of an image.

Those keypoints which are extracted at same location in both images (i.e., binary image and in image mask) are considered, but should resides in extremal region as shown in Figure 18.

The good quality images are described by their invariant features property which should not be effected by any other impediments. The keypoints within extremal regions are important because that describe as a feature of an image.

Some keypoints are consistent in representation but do not appear in extremal region which eventually be rejected and not considered as a focal feature. This might be a drawback of proposed system which can examine later. All extreme points of DoG scale space are located exactly as detected by SIFT. The low contrast and unstable edge points were removed later. At each key point, SIFT computes the scale gradient and direction with respect to the neighborhood. As a reference, SIFT put all calculated values into histogram and summation of these points were used as a gradient of keypoint selection.

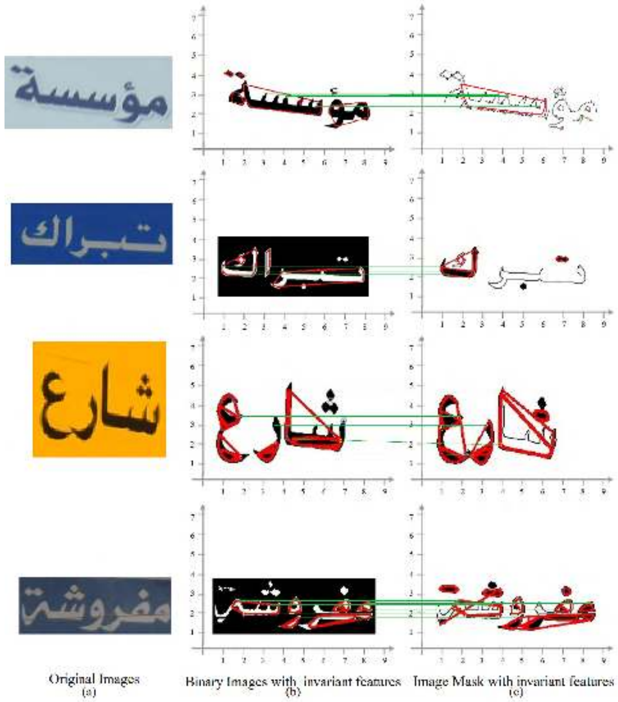

مؤسسة
تبراك
شارع
مفروشة
Original Images
(a)
Binary Images with invariant features
(b)
Image Mask with invariant features
(c)

FIGURE 18. Depiction of extracted invariant features matching.

C. CLASSIFIER FOR TEXT-IN-THE-WILD

The deep learning multidimensional recurrent neural network architecture is proposed as a learning classifier. The recurrent neural network is a suitable architecture where context is important. In Arabic script, the context makes the word meaningful, without context it is difficult to understand the exact word. By keeping this constraint forefront, the potential of context learning approach is investigated on Arabic scene text. Furthermore, recent researches on Arabic like script observed that the said architecture is suitable for complex context learning problems as reported in [3], [6], and [33]–[35]. In basic recurrent neural network architecture, tracing back recent computation makes the history which is maintained by recurrent connection of a neuron. The retained computation in a node would impact in computing the current weight calculation of sequence node. The main constraint in retaining the previous computation is a time lag which varies with subject to given problem. In a situation where problem is large and there is a need to keep all previous calculations intact, in this scenario simple recurrent neural architecture does not support to maintain history for larger input. The Long Short Term Memory (LSTM) networks help to overcome this problem. The LSTM keeps the history as long as it is required, and later forget through its gating mechanism. In LSTM architecture, the hidden neurons were replaced with LSTM memory blocks and it's multiplicative units. The history is manipulated within the memory block by multiplicative units which are responsible to retain or discard the gradient information based on sequence computation requirement at particular point in time. In proposed system, the common keypoints information that appeared in detected extremal regions of binary image and image mask are passed conjunctively and be recorded as a keypoint descriptor. The number of experiments are executed to examine the performance of presented technique.

VI. EXPERIMENTATION

The proposed technique has been evaluated on EASTR-42k dataset having Arabic and English text images. The dataset is splitted into train set, validation set and test set as depicted in Table 5.

TABLE 5. Dataset split for Arabic and English samples.

Dataset Division	Arabic samples	English samples
Train set	7000	3000
Validation set	2000	1200
Test set	1500	1300

As explained earlier, the binary image and image mask are considered for Arabic text analysis. The region was estimated by adapted MSER while invariant features in a region were detected by adapted SIFT as shown in Figure 27.

In English text recognition, window based approach is used for character and word recognition. The images were converted into gray scale and involved pixels in a window were treated as a features. The details about performed experiments on EASTR-42k dataset is mentioned in subsequent section.

A. EVALUATION METRIC

The proposed method is evaluated with similar evaluation metric used by recently reported state-of-the-art work to compute accuracy as reported by Epstein *et al.* [36] and Tian *et al.* [37]. The recognition accuracies on scale invariant binary image and image masks are calculated. In addition, the match between two images is computed through their common detected extremal regions. The extremal regions with intersected keypoints are treated as a features which are trained by classifier for given train set. The accuracy is measured by following equations where T_p is a true positives which means correctly predict the provided samples and F_p is a false positives which means incorrectly predict/ identified the wrong pattern as correct one. Whereas, T_n refers to true negative which correctly predict wrong samples, and F_n represents incorrectly identified correct samples as wrong ones. The relation among T_p,F_p, T_n and F_n is represented in following equations.

$$precision = \frac{\sum T_p}{T_p + F_p} \tag{9}$$

$$recall = \frac{\sum T_p}{T_p + F_n} \tag{10}$$

After calculating precision and recall, the accuracy of learned samples is measured by f-measure as follows,

$$F1 = 2.\frac{precision.recall}{precision + recall} \tag{11}$$

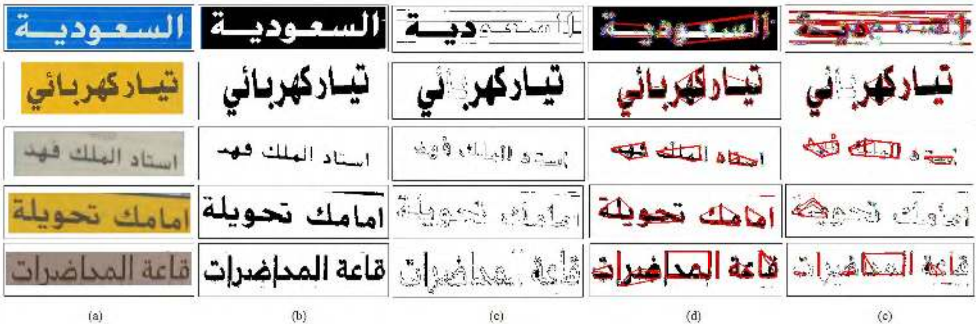


FIGURE 19. (a) Original images. (b) Binary images (c) Image masks (d) Extremal region detected in binary image (e) Extremal region detected in image mask.

B. MDLSTM NETWORK TRAINING FOR ARABIC SCENE TEXT

In MDLSTM, there are as many LSTM memory block as dimensions in an image. At each point in sequence, the network receives external input and its previous own activation along with all dimensions. It is suitable for context learning applications. It has been applied on various research tasks in document image analysis specifically relevant to cursive script as reported in [3]–[5] and [9]. The proposed system is investigated by multidimensional LSTM networks because it maintains contextual information and temporarily correlates the new sequences with previous one. The basic RNN architecture does not retain feature for longer period of time in case of complex input, this problem is regarded as vanishing gradient problem. The aforementioned problem was overcome by introduction of memory blocks instead of hidden neurons. Each memory block comprises memory cell and multiplicative unit. The input is regulated through input gate to memory cell via multiplicative units as indicated in depiction of LSTM classifier in Figure 18. After empirically selected suitable parameters, the segmented Arabic text image is passed to MSER and then invariant features were extracted. In each experiment, as mentioned in Table 8 invariant points are provided to classifier. The invariant points have information about orientation which helps MDLSTM classifier to learn the pattern as it appears but with the coordinate values. The coordinate values of invariant points provide primitive feature information which is further manipulated for learning purpose. As input is complex and cursive in nature, 5 hidden layers are used comprising 20, 40, 60, 80 and 100 LSTM memory blocks in each layer. As architecture of RNN suggests, the hidden layers are fully connected. The hidden blocks are fed to feed forward network having tanh summation units for the purpose to activate cell. All these hidden layer processing collapsed into one dimensional sequence and later Connectionist Temporal Classification (CTC) labels the learned patterns.

The network's optimal performance can be obtained by careful consideration and selection of effective parameters. Figure 20 represents the best learning curve obtained during training on character dataset. The red line represents the best point while learning the given samples. After the red line the difference between training and validation increases which ultimately culminate the training after 450 epochs.

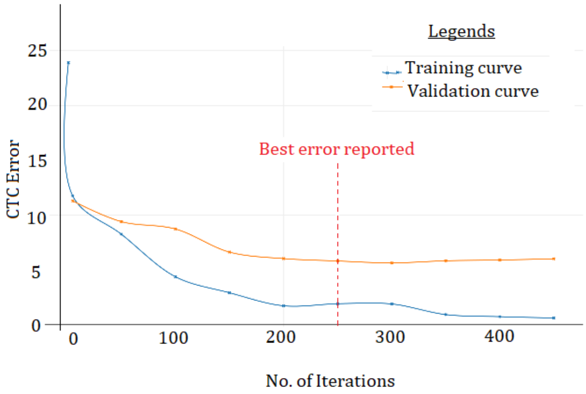


FIGURE 20. Learning curve with 100 LSTM memory units.

The number of parameters specific to recurrent neural network like input block, learning rate, and number of units in LSTM memory block. With reference to parameters the input block defines as 4 × 4, it means the invariant points resides in 4 × 4 block would be given as an input. In other words, the hidden block size means that the features are collected into 4 × 4 block size. The preliminary experiments guided towards correct or optimal selection of parameters. The selected parameters for training MDLSTM on Arabic scene text is presented in Table 6. The training stops at a point where no improvement observed during validation.

TABLE 6. Parameter selection during training.

Parameters	Values
Input block size	4 × 4
Hidden block size	4 × 4
Hidden memory units	20, 40, 60, 80, 100
Learning rate	1 x 10^{-4}
Momentum	0.9
Total network weight for exp 1	2,80,273
Total network weight for exp 2	2,10,028
Total network weight for exp 3	93,724

TABLE 7. Time comparison per iteration on various LSTM memory blocks.

Hidden Layer sizes	20	40	60	80	100
Time per iteration (seconds)	18.3	29.7	44.1	60.5	73.0

The experiments are conducted on different LSTM hidden memory size as mentioned in Table 7, which represents the number of hidden memory units and consumed time to perform training on given samples. In order to achieve optimal accuracy, the hidden layer units having size of 20, 40, 60, 80, *and* 100 are considered for a purpose to evaluate the potential of proposed system.

TABLE 8. Accuracies reported on trained data.

Exp-Variations	No. of SIFTs	No. of Hidden Layers (%)				
		20	40	60	80	100
Binary+SIFT+MSER	2,80,273	60.12	63.70	71.53	79.17	90.44
Mask+SIFT+MSER	2,10,028	61.78	63.42	74.16	83.29	92.57
Intersection	93,724	69.72	71.89	79.24	87.61	94.50

In Table 8, the training accuracy is reported by considering various LSTM memory blocks. The number of SIFT points were also counted against each combination of experiments. Here, by looking at the table, there is an assumption that if provide less number of interested common SIFT features then it may result in achieving good accuracy.

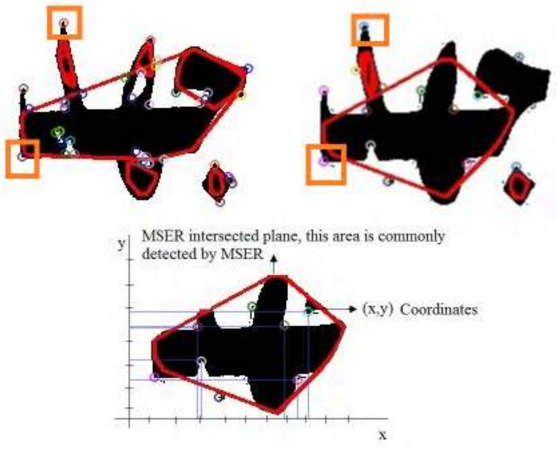

y
MSER intersected plane, this area is commonly detected by MSER
(x,y) Coordinates
x

FIGURE 21. Conjoined MSER region mapped on spatial domain.

This will provide an efficient solution because every keypoint extracted by SIFT is not important to consider rather the keypoint appeared in extremal region is more relevant for training purpose as shows in Figure 21. As represented in figure, the common MSER region is mapped on x-y coordinates. The orange box indicates that SIFT features were detected on same place but that place has not been detected by MSER so, these points which could include for classification would go in vain as the common points lies in detected extremal region re-considered.

TABLE 9. Accuracies reported during training of word data on ASTR-27k dataset.

Parameters	V_1	V_2	V_3
Number of SIFTs	2, 80, 273	2, 10, 028	93, 724
t_p	47.24%	68.71%	71.24%
f_p	10.78%	6.16%	4.71%
t_n	23.47%	17.59%	15.12%
f_n	18.51%	7.54%	8.93%
precision	0.81	0.92	0.94
Recall	0.72	0.90	0.89
F-measure	0.76	0.91	0.94

C. PERFORMED EXPERIMENTS ON ARABIC SCENE TEXT

As mentioned earlier, that under discussion experimental study is carried out on 1500 scene text images where text has been segmented into words and invariant features were extracted. For third experiment the assumption is that all extracted keypoints features are not important to examine therefore intersection of keypoints are considered which are detected in two different images and pass the coordinate values of specific regions to classifier.

Figure 22 represents the training curves observed on various LSTM memory blocks during word recognition. The network shows overfitting trend when memory blocks were 20 in size as shown in figure. The training and validation curves were also mapped in this figure on 40, 60, 80 and 100 LSTM memory blocks. The best accuracy was reported when LSTM memory blocks was 100. The learning curves represent trade-off between training and validation set. The accuracy is reported with respect to characters. In proposed dataset each character in Arabic is appeared in different styles and in various orientation, so the invariant features extraction approach is suitable choice for them to classify. Figure 23 shows bad sample images exist in acquired dataset which impact the final accuracy.

As mentioned in Table 8, The experimental analysis is performed by considering three variations on Arabic word recognition in natural images. In other experiments, only common invariant points in extremal regions are considered for character and text line recognition as summarized in Table 12.

In word recognition, the process of extracting SIFT features and detecting extremal regions are common but applied it on two different images i.e., binary and image mask, whereas the third variation is considered the intersected points as features and classify them. The first variation which used the invariant feature extraction and extremal regions detection approach on image mask, is named as V_1. Similarly, other experimental variation with binary image and intersected points are referred as V_2 and V_3 respectively. The number of SIFT features were counted, then compute accuracy by precision, recall and F-measure as detailed in Table 10. There are some samples which represented missing or blank image characters after pre-processing. Such samples might influence with illumination or orientation factors as shown in Figure 24.

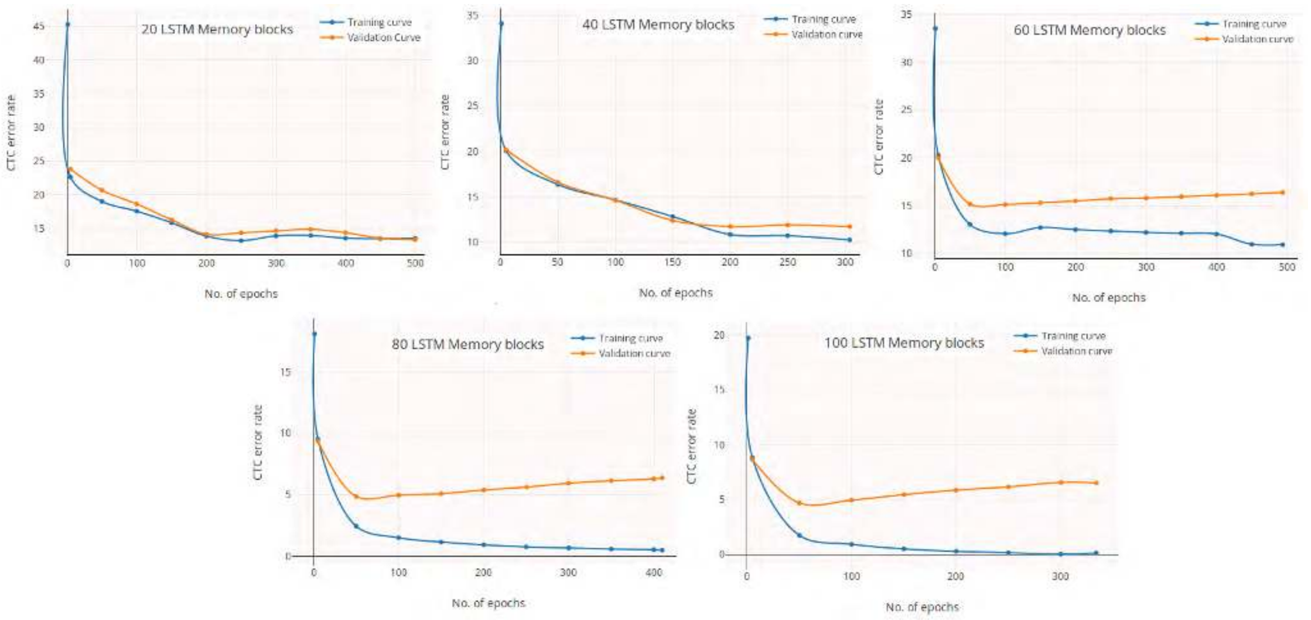


FIGURE 22. Training curves on various LSTM memory block.

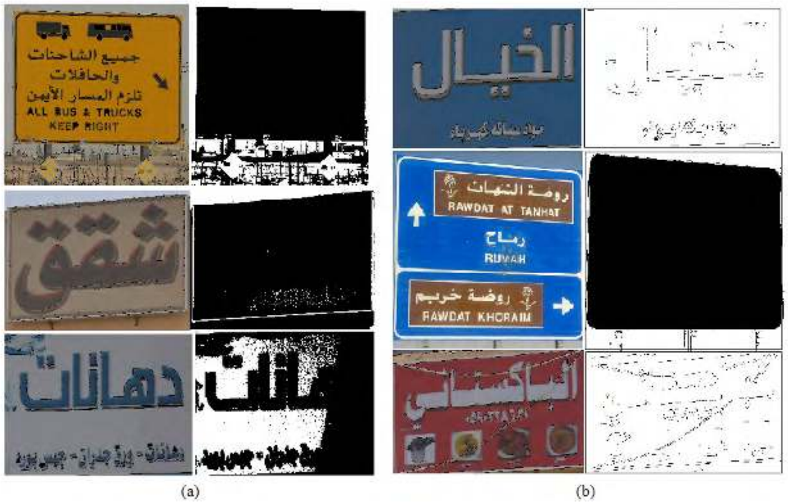


FIGURE 23. Examples of some bad images.(a) Misclassified text region in binary images. (b) Text detection is ambiguous in some image masks.

D. PERFORMED EXPERIMENTS ON ENGLISH SCENE TEXT

The multilingual nature of acquired images impel to prepare dataset for English. The heterogeneous approaches have been experimented on English scene text detection and recognition as represented in [36] and [38]–[40].

In this paper the detail about performed experiments on English data samples as part of EASTR-42k dataset is also provided. The conducted experiments examined English characters and words appeared in EASTR-42k. As depicted the overall process in Figure 25, the evaluation was performed on gray scale image by normalizing the x-height to 40 and 60. After rescaling, the image was readjusted by aspect ratio. The window based features were provided to MDLSTM classifier in a same manner as mentioned in [6] for Latin character recognition. The good accuracy is accomplished in terms of character and word recognition as summarized in Table 10.

TABLE 10. Evaluation results on EASTR-42k English samples.

x-height	Char Accuracy	Word Accuracy
40	96.34	97.14
60	95.19	97.31

As English language does not fall into cursive script family, so the recognition of English scene character provides very encouraging results especially in evaluation of EASTR-42k dataset. The comparison of proposed work with recently

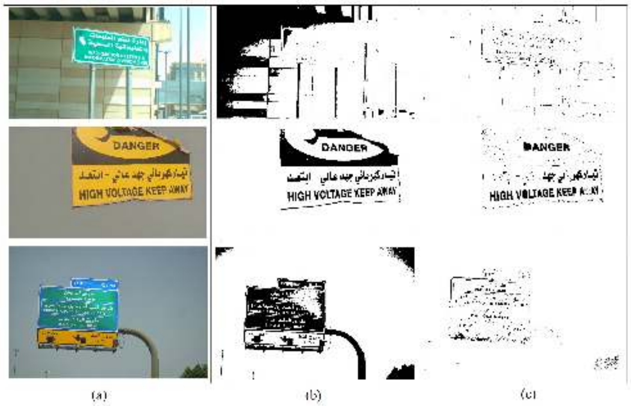


FIGURE 24. Sample images where scene text was influenced by light or orientation.

proposed algorithms designed for Arabic and English scene text recognition is presented in next section.

VII. PERFORMANCE COMPARISON

This section presents comparison of recent research on English and Arabic scene text recognition by keeping focus on performance aspect. The comparison of English scene text research presented in recent years is provided on the basis of performance reported using deep learning methods as summarized in Table11. The performance of several presented work on English during the last five years are summarized in this table. The deep learning architectures are proposed in recent years to boost the performance of learning methods having large corpus. The deep neural networks (DNN) [44], [51], attention modeling [47], convolutional neural networks (CNN) [41]–[43], [52] are the prime deep learning architecture which are adapted in relevance to provided samples for training purpose. The obtained performance is measured by precision, recall and f-measure as depicted in table.

The work on deep learning convolutional network based isolated Arabic scene character recognition is presented by Ahmed *et al.* [3]. They evaluated their presented method on ASTR dataset and divided the taken samples into trainset and testset. They identified 27 classes whereas as each class is rescaled into 50 × 50 size. They consider five different orientations with respect to various angles. They consider 2450 character images for training while the performance of learned network is evaluated on 250 images. Their proposed architecture used 2 convolutional layers followed by fully connected layer. In comparison to other available work in Arabic scene text analysis, they reported 0.15% error rate.

Figure 26 represents sample images in EASTR-42k image dataset. As shown in figure, the images are complex representing challenging text in various font styles, sizes, color and in different angles having complex background. ASTR dataset is a subset of proposed dataset which only contains Arabic text samples. As observed from the figure that samples were acquired in an uncontrolled environment where text has rendered to numerous challenges.

The robust algorithm for detection of Arabic video text is presented by Halima *et al.* [54]. In this paper, they proposed Laplacian operator used for text detection. They identified the candidate region by Laplacian operator in frequency domain whereas, the edges were detected by projection profile method. They measure the performance by calculating the precision 0.96% and recall 0.95%.

One of the latest work proposed by Jain *et al.* [55], they demonstrated sub-sampling approach by deep learning classifier to evaluate the screen rendered Arabic text. They used convolutional network to segment the input sample against target labels and learn the contextual dependencies among part of segmented sample. They evaluated their results on freely available synthetic Arabic scene text datasets named ACTIV [56] and ALIF [57]. The third dataset they prepared by downloading available Arabic scene text on Google Image. They reported 98.17% accuracy on ALIF dataset while on ACTIV dataset they achieved 97.44% accuracy. A comprehensive model for Arabic text, detection, localization, extraction and recognition is proposed by Halima *et al.* [59]. The text has been detected by using four different techniques which are, connected component methods, texture classification methods, edge detection methods and correlation based methods. They normalized the samples by setting x-height to 26 pixels. They considered dot over the characters as one feature while all characters do not have dots that's why main body of character also considered like projection feature, transition feature and occlusion extraction are main features relevant to single character. They used supervised learning k-nearest neighbor algorithm to learn the patterns. They measure recall, precision and f-score on their proposed dataset.

Another method for text localization of Farsi script is presented by Darab and Rahmati [60]. They detected candidate text by considering edge and color information. The features were extracted by wavelet coefficient histogram features. The SVM is used to classify the text and non-textual pattern. There is not benchmark dataset for Farsi is available, therefore they also suggested new dataset for Farsi scene text. They disintegrate their experiments by localization techniques and feature comparison. The performance was evaluated by calculating precision, recall and f-measure on localization and feature techniques separately. They evaluate their proposed method on HOG, Wavelet coefficient histogram and combination of both and reported precision 76.0%, 62.6%, 80.8% whereas, the recall was 76.0%, 71.5% and 29.4% respectively. The f-score for earlier mentioned experiments were 76.0%, 83.3% and 86.5% respectively. The proposed work measure character, word and line recognition rate of acquired dataset images and compare its results with recently proposed work in Table 12.

The camera captured scene text recognition competition was organized by ICDAR in 2015 edition. The challenge 2 of ICDAR 2015 was focused on camera captured focused scene text as explained by Zhou *et al.* [61]. In addition to previous tasks, they introduced End-to-End system

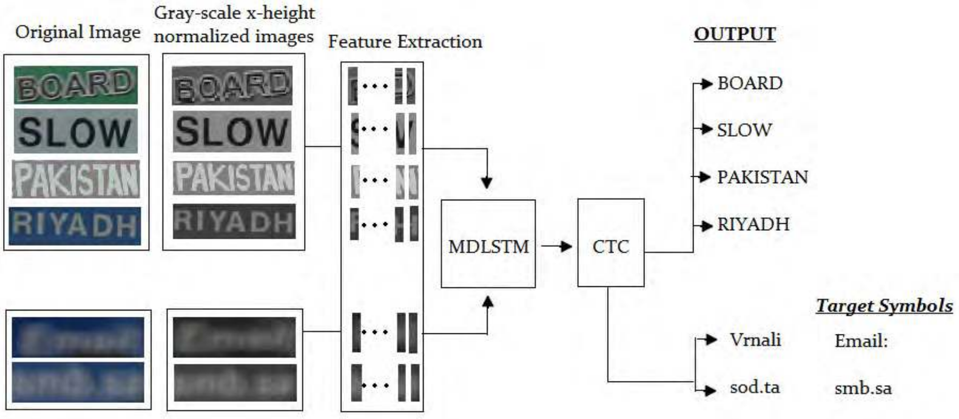


FIGURE 25. Depiction of proposed methodology and expected output for good and bad images in English scene text recognition EASTR-42k dataset.

TABLE 11. Comparison of deep learning based English research work.

Study	Year	Methodology	DB name/size	Precision	Recall	F-score
Proposed Algorithm	NA	Deep-MDLSTM	ESTR-15k	94.1	89.5	97.52
Christian et al [44]	2018	Deep Neural Netwrok (DNN)	SVHN [45], FSNS [46]	NR	NR	95.0, 78.0,
Yuting et al [47]	2018	Attention Modelng with double supervised network	IIIT5K, ICDAR 2013, SVT	NR	NR	88.6, 92.3, 84.1
Fei et al [42]	2017	Sliding convolutional models	IIIT5k, SVT, ICDAR03/13, TRW15	NR	NR	98.9, 95.1, 97.7, 89.0
Christian et al [48]	2017	Semi-supervised DNN	ICDAR2013, SVT, IIIT5k	NR	NR	90.3, 79.8 and 86.0
Yingying et al [41]	2017	Rotational CNN	ICDAR2013	93.55	82.59	87.73
Yingying et al [41]	2017	Rotational CNN	ICDAR2015	85.62	79.68	82.54
Wenhao et al [49]	2017	Deep Direct Regression	ICDAR2015	82.0	80.0	81.0
Wenhao et al [49]	2017	Deep Direct Regression	MSRA-TD500	77.0	70.0	74.0
Yuliang et al [50]	2017	DMPnets	ICDAR2015	73.23	68.22	70.64
Yuliang et al [51]	2016	Single-DNN	ICDAR2011	88.0	82.0	85.0
Yuliang et al [51]	2016	Single-DNN	ICDAR2013	88.0	83.0	85.0
Yuliang et al [43]	2016	CNN	ICDAR2013	88.0	83.0	85.0
Zhuoyao et al [52]	2016	Deep-CNN	ICDAR2011	85.0	81.0	83.0
Zhuoyao et al [52]	2016	Deep-CNN	ICDAR2013	87.0	83.0	85.0
Anupama et al [53]	2014	Deep Belief Networks (DBN)	Char74k	NR	NR	84.04

*NR = Not Reported

TABLE 12. Comparison of Arabic scene text recognition results with recently proposed Arabic dataset.

Study	Year	Methods	Dataset	CRR	WRR	LRR
Proposed Algorithm	2018	SIFT+MSER+MDLSTM	ASTR-27k	96.32	94.01	75.20
Oussama et al [2]	2018	MDLSTM	AcTiV 2.0	94.0	NR	62.0
Ahmed et al [3]	2017	ConvNets	EASTR	85.3	NR	NR
Yousfi et al [57]	2015	ConvNets+BLSTM	ALIF_Test1	94.36	71.26	55.03
Yousfi et al [57]	2015	CovNets+BLSTM	ALIF_Test2	90.71	65.67	44.90
Tounsi et al [58]	2015	Sparse Coding	ARASTEC(Char74k-15)	73.1	NR	NR
Tounsi et al [58]	2015	Sparse Coding	ICDAR03-15	75.3	NR	NR

*NR = Not Reported

performance tasks in ICDAR 2015 competition. The ground truth is defined at word level. As reported in their paper, the most of presented techniques used Maximally Stable Extremal Regions (MSER) for text localization whereas, top performing methods in competition used commercial OCRs for recognition purpose. In comparison to previous years

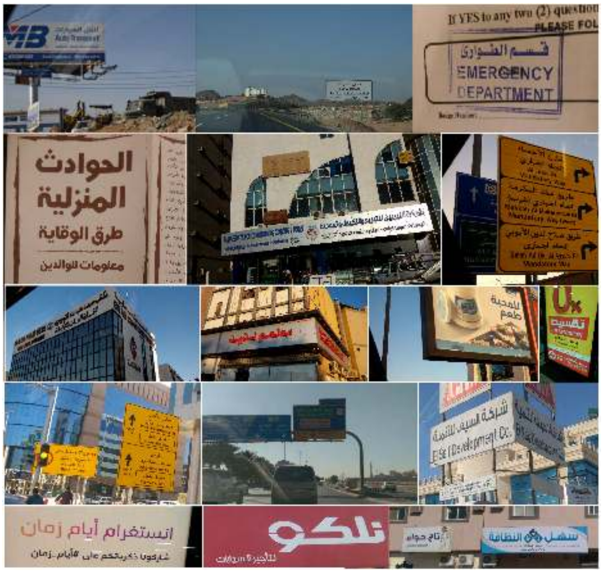


FIGURE 26. Challenging text images in EASTR-42k dataset.

competition, ICDAR 2015 competition marked significant number of increased researchers that depicts their interest as far as scene text analysis is concerned. The noteworthy increased recognition rate is observed by using submitted methods. The dataset has been distinguished on the basis of contextual complexities of vocabulary words. The eight submitted techniques were experimented as described in their paper and summarized in Table 13.

TABLE 13. Comparison of ICDAR 2015 Word Recognition Results with proposed method.

Method	F-measure (%)
Ours	97.52
VGGMaxBBNet	86.18
Stradivision-1	81.28
Baseline (Text Spotter)	77.02
Deep2Text-I	45.1
MSER-MRF	71.13
Beam Search CUNI	63.2
Baseline (OpenCV + Tesseract)	59.47
Beam Search CUNI + S	26.38

ICDAR 2017 robust reading competition was held in November 2017 for its fifth edition. The organizers of competition offered numerous challenges among them multilingual scene text detection and script identification challenge was offered. Although the competition has closed in March 2018, their results has not published yet. Therefore, yielded results can not be compare with them.

In presented work, the problems are highlighted and sketched the solution to address complications that exists in cursive text recognition in natural images. A comprehensive benchmark dataset for Arabic scene text recognition is also prepared which is integrated into textlines, words and characters. The multilingual appearance of captured samples allow to prepare dataset for Latin script along with Arabic. In this way, there is huge collection of various characters and words representation that appeared without any specific font style presented. The dataset for Arabic numerals also presented. The Arabic samples represented in ASTR while English samples included in ESTR. This paper presented hybrid feature extraction approach that considers the relevant regions and invariant point occurred in that region. The validation and verification of proposed dataset is performed by MDLSTM due to its strong sequence learning ability. The reported experimental analysis is on ASTR dataset. The accuracy is computed by recall and precision. The presented work achieved state-of-the-art results in comparison to recently reported work and the work presented in ICDAR competition. This work considered as a benchmark effort because of scarcity of work in relevance to Arabic script.

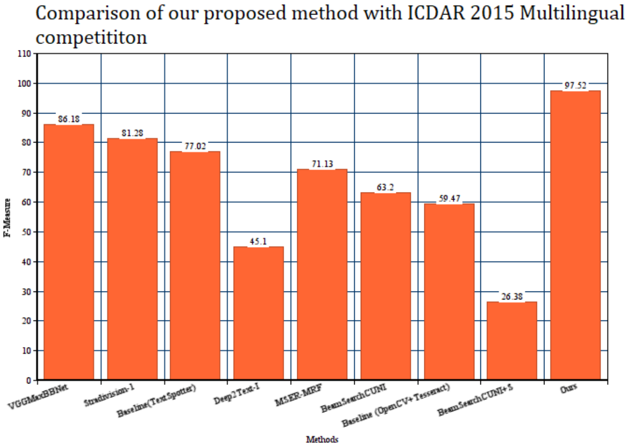


FIGURE 27. ICDAR 2015 submitted approaches comparison with presented methodology.

VIII. CONCLUSION

This paper presented a novel hybrid method that has focused on relevant features for classification. The performance of proposed system was evaluated on three experimental variations to investigate the behavior of MDLSTM network on proposed method. In each experiment, number of invariant points in all images were counted. For Arabic scene text images, the extremal regions were first detected and invariant feature were extracted from binary images and image masks. The more to concentrate on irrelevant invariant features resulted as worse performance. Therefore, in the presented work the extremal regions were detected and invariant features were also accounted by SIFT approach in conjoined region. As observed from experiments that invariant feature extraction approach only provide better result if there is a focus on specific area for invariant features. The feature intersection of both images (i.e., binary image and image mask) as detected in extremal regions displayed benchmark results i.e., error rates of 5.98% on Arabic and 2.48% on English are computed. Arabic scene text dataset is also presented, which covers variety of Arabic scene text images appeared in various illumination and effected by unconstrained environment. The proposed dataset also focuses on Latin script as most of the Arabic text available in English at a same time. As proposed work is a novel effort specifically in Arabic text recognition in natural images with large corpus. Therefore, comparison of presented technique was performed with recently proposed techniques by considering other cursive and non-cursive scene text.

In future, there is a plan to utilize the strength of instance learning capability of convolutional neural network with context learning long short term memory networks. Their combination assumed to provide good results on unconstrained cursive scene text recognition.

REFERENCES

[1] A. Zhu, R. Gao, and S. Uchida, "Could scene context be beneficial for scene text detection?" *Pattern Recognit.*, vol. 58, pp. 204–215, Oct. 2016.

[2] O. Zayene, S. M. Touj, J. Hennebert, R. Ingold, and N. E. Ben Amara, "Open datasets and tools for arabic text detection and recognition in news video frames," *J. Imag.*, vol. 4, no. 2, p. 32, 2018. [Online]. Available: http://dblp.uni-trier.de/db/journals/jimaging/jimaging4.html#ZayeneTHIA18

[3] S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, "Deep learning based isolated arabic scene character recognition," in *Proc. 1st IEEE Workshop Arabic Script Anal. Recognit.*, Apr. 2017, pp. 46–51.

[4] L. Li, S. Yu, L. Zhong, and X. Li, "Multilingual text detection with nonlinear neural network," *Math. Problems Eng.*, vol. 2015, Sep. 2015, Art. no. 431608, doi: 10.1155/2015/431608.

[5] S. Naz, A. I. Umar, R. Ahmed, M. I. Razzak, S. F. Rashid, and F. Shafait, "Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks," *SpringerPlus*, vol. 5, no. 1, p. 2010. [Online]. Available: http://www.springerplus.com/content/5/1/201

[6] S. B. Ahmed, S. Naz, M. I. Razzak, S. F. Rashid, M. Z. Afzal, and M. Thomas Breuel, "Evaluation of cursive and non-cursive scripts using recurrent neural networks," *Neural Comput. Appl.*, vol. 27, no. 3, pp. 603–613, 2016.

[7] H. A. Elnemr, "Combining SURF and MSER along with color features for image retrieval system based on bag of visual words," in *J. Comput. Sci.*, vol. 12, no. 4, pp. 213–222, 2016. [Online]. Available: http://dblp.uni-trier.de/db/journals/jcsci/jcsci12.html#Elnemr16

[8] S. B. Ahmed, S. Naz, M. I. Razzak, R. Yusof, and R. Yusof, "Arabic cursive text recognition from natural scene images," *Appl. Sci.*, vol. 9, no. 2, p. 236, 2019.

[9] S. B. Ahmed, S. Naz, M. I. Razzak, R. Yusof, and T. M. Breuel, "Balinese character recognition using bidirectional LSTM classifier," in *Proc. Adv. Mach. Learn. Signal Process. (MALSIP)*, 2015, pp. 201–211, doi: 10.1007/978-3-319-32213-1_18.

[10] S. B. Ahmed, S. Naz, M. I. Razzak, S. F. Rashid, M. Z. Afzal, and T. M. Breuel, "Evaluation of cursive and non-cursive scripts using recurrent neural networks," *Neural Comput. Appl.*, vol. 27, pp. 603–613, Apr. 2016.

[11] S. B. Ahmed, S. Naz, S. Swati, and M. I. Razzak, "Handwritten urdu character recognition using 1-dimensional BLSTM classifier," *Neural Comput Appl.*, vol. 13, pp. 1–9, 2017, doi: 10.1007/s00521-017-3146-x.

[12] Y.-F. Pan, X. Hou, and C.-L. Liu, "Text localization in natural scene images based on conditional random field," in *Proc. ICDAR*, 2009, pp. 6–10, doi: 10.1109/ICDAR.2009.9.

[13] L. Gómez and D. Karatzas, "MSER-based real-time text detection and tracking," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 3110–3115. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6966883

[14] L. Neumann and J. Matas, "Efficient scene text localization and recognition with local character refinement," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2015, pp. 746–750. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1504.html#NeumannM15

[15] A. Graves, "Supervised sequence labelling," in *Supervised Sequence Labelling With Recurrent Neural Networks*, vol. 385. Berlin, Germany: Springer, 2012, pp. 1–131, doi: 10.1007/978-3-642-24797-2.

[16] M. Tounsi, I. Moalla, and A. M. Alimi, "ARASTI: A database for Arabic scene text recognition," in *Proc. ASAR*, 2017, pp. 140–144. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8054539

[17] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Computer Vision—ACCV* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2010, pp. 770–783.

[18] J. Fabrizio, B. Marcotegui, and M. Cord, "Text segmentation in natural scenes using toggle-mapping," in *Proc. IEEE ICIP*, Nov. 2009, pp. 2373–2376. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5403221

[19] O. R. Chum and J. Matas, "Optimal randomized RANSAC," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1472–1482, Aug. 2008.

[20] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. CVPR*, Jun. 2012, pp. 1083–1090. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6235193

[21] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018. [Online]. Available: http://arxiv.org/abs/1703.01086

[22] Y. Tang and X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2276–2288, Sep. 2018. [Online]. Available: http://dblp.uni-trier.de/db/journals/tmm/tmm20.html#TangW18

[23] S. Tian, X.-C. Yin, Y. Su, and H.-W. Hao, "A unified framework for tracking based text detection and recognition from Web videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 542–554, Mar. 2018, doi: 10.1109/TPAMI.2017.2692763.

[24] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Apr. 2018. [Online]. Available: http://arxiv.org/abs/1801.02765

[25] Y. Tang and X. Wu, "Scene text detection and segmentation based on cascaded convolution neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1509–1520, Mar. 2017. [Online]. Available: http://dblp.uni-trier.de/db/journals/tip/tip26.html#TangW17

[26] X. Ren, Y. Zhou, Z. Huang, J. Sun, X. Yang, and K. Chen, "A novel text structure feature extractor for chinese scene text detection and recognition," *IEEE Access*, vol. 5, pp. 3193–3204, 2017. [Online]. Available: http://dblp.uni-trier.de/db/journals/access/access5.html#RenZHSYC17

[27] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. ICDAR*, 2003, pp. 682–687.

[28] S. M. Lucas, "ICDAR 2005 text locating competition results," in *Proc. ICDAR*, 2005, pp. 80–84, doi: 10.1109/ICDAR.2005.231.

[29] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. ICDAR*, 2011, pp. 1491–1496. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6065245

[30] D. Kumar, M. N. A. Prasad, and A. G. Ramakrishnan, "Multi-script robust reading competition in ICDAR 2013," in *Proc. 4th Int. Workshop Multilingual OCR*, Washington, DC, USA, Aug. 2013, pp. 14-1–14-5, doi: 10.1145/2505377.

[31] J. van Beusekom, F. Shafait, and T. M. Breuel, "Combined orientation and skew detection using geometric text-line modeling," *Int. J. Document Anal. Recognit.*, vol. 13, no. 2, pp. 79–92, 2010. [Online]. Available: http://dblp.uni-trier.de/db/journals/ijdar/ijdar13.html#BeusekomSB10

[32] M. Ben Halima, H. Karray, A. M. Alimi, and A. F. Vila. (2012). "NF-SAVO: Neuro-fuzzy system for arabic video OCR." [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1211.html#abs-1211-2150 and https://arxiv.org/abs/1211.2150

[33] S. Naz *et al.*, "Urdu Nastaliq recognition using convolutional–recursive deep learning," in *Neurocomputing*, vol. 243, pp. 80–87, Jun. 2017, doi: 10.1016/j.neucom.2017.02.081.

[34] S. Naz *et al.*, "Offline cursive Urdu–Nastaliq script recognition using multidimensional recurrent neural networks," *Neurocomputing*, vol. 177, pp. 228–241, Feb. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121501749X, doi: 10.1016/j.neucom.2015.11.030.

[35] S. Bin Ahmed, S. Naz, M. I. Razzak, and R. Yusof, "Cursive scene text analysis by deep convolutional linear pyramids," in *Proc. 25th Int. Conf. Neural Inf. Process. (ICONIP)*, 2018, pp. 307–318. [Online]. Available: http://arxiv.org/abs/1809.10792

[36] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. CVPR*, 2010, pp. 2963–2970. [Online]. Available: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2010.html#EpshteinOW10

[37] S. Tian *et al.*, "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognit.*, vol. 51, pp. 125–134, Mar. 2016. [Online]. Available: http://dblp.uni-trier.de/db/journals/pr/pr51.html#TianBLSWWLT16

[38] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and benchmark for text detection and recognition in natural images," in *Proc. CVPR*, 2016. [Online]. Available: http://arxiv.org/abs/1601.07140

[39] C. Yi and Y. Tian, "Text extraction from scene images by character appearance and structure modeling," *Comput. Vis. Image Understand.*, vol. 117, no. 2, pp. 182–194, Feb. 2013, doi: 10.1016/j.cviu.2012.11.002.

[40] Q. Yang, M. Cheng, W. Zhou, Y. Chen, M. Qiu, and W. Lin, "IncepText: A new inception-text module with deformable PSROI pooling for multi-oriented scene text detection," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Stockholm, Sweden, Jul. 2018, pp. 1071–1077. [Online]. Available: http://www.ijcai.org/proceedings/2018/

[41] Y. Jiang *et al.* (2017). "R2CNN: Rotational region CNN for orientation robust scene text detection." [Online]. Available: https://arxiv.org/abs/1706.09579

[42] F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. Liu. (2017). "Scene text recognition with sliding convolutional character models." [Online]. Available: https://arxiv.org/abs/1709.01727

[43] T. He, W. Huang, Y. Qiao, and J. Yao. (2016). "Accurate text localization in natural image with cascaded convolutional text network." [Online]. Available: https://arxiv.org/abs/1603.09423

[44] C. Bartz, H. Yang, and C. Meinel. (2017). "SEE: Towards semi-supervised end-to-end scene text recognition." [Online]. Available: https://arxiv.org/abs/1712.05404

[45] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011, pp. 1–9.

[46] R. Smith *et al.*, "End-to-end interpretation of the french street name signs dataset," in *Proc. Comput. Vis. Workshops (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 411–426.

[47] Y. Gao, Z. Huang, and Y. Dai. (2018). "Double supervised network with attention mechanism for scene text recognition." [Online]. Available: https://arxiv.org/abs/1808.00677

[48] C. Bartz, H. Yang, and C. Meinel. (2017). "STN-OCR: A single neural network for text detection and text recognition." [Online]. Available: https://arxiv.org/abs/1707.08831

[49] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proc. ICCV*, 2017, pp. 745–753. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8234942

[50] Y. Liu and L. Jin. (2017). "Deep matching prior network: Toward tighter multi-oriented text detection." [Online]. Available: https://arxiv.org/abs/1703.01425

[51] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. (2016). "TextBoxes: A fast text detector with a single deep neural network." [Online]. Available: https://arxiv.org/abs/1611.06779

[52] Z. Zhong, L. Jin, S. Zhang, and Z. Feng. (2016). "DeepText: A unified framework for text proposal generation and text detection in natural images." [Online]. Available: https://arxiv.org/abs/1605.07314

[53] A. Ray, S. Rajeswar, and S. Chaudhury, "Scene text analysis using deep belief networks," in *Proc. Indian Conf. Computer Vis., Graph. Image Process. (ICVGIP)*, 2014, Art. no. 71. [Online]. Available: http://dl.acm.org/citation.cfm?id=2683483,"

[54] M. Ben Halima, H. Karray, and A. M. Alimi, "Arabic text recognition in video sequences," *Int. J. Comput. Linguistics Res.*, pp. 603–608, Aug. 2013. [Online]. Available: http://arxiv.org/abs/1308.3243

[55] M. Jain, M. Mathew, and C. V. Jawahar, "Unconstrained scene text and video text recognition for Arabic script," in *Proc. 1st Workshop Arabic Script Anal. Recognit.*, 2017, pp. 26–30. [Online]. Available: http://arxiv.org/abs/1704.06821

[56] O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, and N. E. B. Amara, "A dataset for Arabic text detection, tracking and recognition in news videos- AcTiV," in *Proc. ICDAR*, 2015, pp. 996–1000. [Online]. Available: http://dblp.uni-trier.de/db/conf/icdar/icdar2015.html#ZayeneHTIA15

[57] S. Yousfi, S.-A. Berrani, and C. Garcia, "ALIF: A dataset for Arabic embedded text recognition in TV broadcast," in *Proc. ICDAR*, 2015, pp. 1221–1225. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7321714

[58] M. Tounsi, I. Moalla, A. M. Alimi, and F. Lebouregois, "Arabic characters recognition in natural scenes using sparse coding for feature representations," in *Proc. ICDAR*, 2015, pp. 1036–1040. [Online]. Available: https://ieeexplore.ieee.org/document/7333919/

[59] M. Ben Halima, H. Karray, and A. M. Alimi, "A comprehensive method for arabic video text detection, localization, extraction and recognition," in *Advances in Multimedia Information Processing—PCM* (Lecture Notes in Computer Science), vol. 6298. Berlin, Germany: Springer, 2010, pp. 648–659. [Online]. Available: http://dblp.uni-trier.de/db/conf/pcm/pcm2010-2.html#HalimaKA10

[60] M. Darab and M. Rahmati, "A hybrid approach to localize farsi text in natural scene images," in *Proc. 3rd Int. Neural Netw. Soc. Winter Conf. (INNS-WC)*, Bangkok, Thailand, vol. 13, Oct. 2012, pp. 171–184. [Online]. Available: http://www.sciencedirect.com/science/journal/18770509/13

[61] X. Zhou, S. Zhou, C. Yao, Z. Cao, and Q. Yin. (2015). "ICDAR 2015 text reading in the wild competition." [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1506.html#ZhouZYCY15

SAAD BIN AHMED received the M.Sc. degree in intelligent systems from Technische Universitaet, Kaiserslautern, Germany. He is currently pursuing the Ph.D. degree in intelligent systems with Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia. He was also appointed as a Lecturer with the COMSATS Institute of Information Technology, Abottabad, Pakistan. He is a Lecturer with King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia. He is also with the Center of Artificial Intelligence and Robotics, Malaysia–Japan Institute of Information Technology, Universiti Teknologi Malaysia. He served as a Research Assistant with the Image Understanding and Pattern Recognition Research Group, University of Technology, Kaiserslautern. He has authored more than 20 papers in impact factor journals, conferences, and book chapters. His areas of interests include document image analysis, machine learning, computer vision, and optical character recognition. He has been involved in the field of image analysis and pattern recognition field, since 15 years, and has been involved in various pioneer researches like collection of handwritten Urdu data and used it for Urdu character recognition. He is also providing his expertise in capturing Arabic scene text images and performing research on collected samples by machine learning and pattern recognition techniques.

SAEEDA NAZ received the B.S. degree from the University of Peshawar, Peshawar, Pakistan, in 2006, the M.S. degree in computer science from the COMSATS Institute of Information Technology, Pakistan, in 2012, and the Ph.D. degree (Hons.) in computer science from the Department of Information Technology, Hazara University, Mansehra, Pakistan, in 2016. She is an Assistant Professor and the Head of the Computer Science Department, GGPGC No.1, Higher Education Department of Government of Khyber-Pakhtunkhwa, Abbottabad, Pakistan, since 2008. She has published two book chapters and more than 30 papers in peer-reviewed national and international conferences and journals. Her areas of interests include optical character recognition, pattern recognition, machine learning, medical imaging, and natural language processing.

MUHAMMAD IMRAN RAZZAK was an Associate Professor of health informatics with the College of Public Health and Health Informatics. He is currently a Senior Researcher with the University of Technology, Sydney. He has published more than 70 papers in well reputed journals and conferences. He holds research grants of more than $1.3 million. He has authored more than 60 papers in well reputed journals and conferences. He is an Inventor of one patent. His areas of research include machine learning and health informatics. He has developed and delivered several research projects successfully. He was a recipient of the Young Researcher 2015 NGHA, Saudi Arabia, based on his research contributions, and the Best Researcher during his stay with CoEIA.

RUBIYAH BTE YUSOF received the B.Sc. degree (Hons.) in electrical and electronics engineering from the University of Loughborough, U.K., in 1983, the master's degree in control systems from the Cranfield Institute of Technology, U.K., in 1986, and the Ph.D. degree in control systems from the University of Tokushima, Japan, in 1994. Throughout her career as a Senior Lecturer and a Researcher with Universiti Teknologi Malaysia (UTM), she has been acknowledged for her many contributions in artificial intelligence, process control, and instrumentation design. She was the Director of CAIRO. She is currently the Dean of the Malaysia–Japan Institute of Technology, UTM, Kuala Lumpur. She has authored the book *Neuro-Control and Its Applications* (Springer Verlag, 1995) which was translated to Russian, in 2001. She is recognized for her work in biometrics systems, such as KenalMuka (face recognition systems) and signature verification systems which received both national and international awards. She is a member of the AI Society, Malaysia, the Instrumentation and Control Society, Malaysia, and the Institute of Electrical and Electronics Engineers, Malaysia.