Unconstrained Arabic Scene Text Analysis using Concurrent Invariant Points

Saad Bin Ahmed*†, Saeeda Naz‡, Imran Razzak§ and Mukesh Prasad¶
*Center for Artificial Intelligence and Robotics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
† Health Informatics, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
‡Govt. Girls Postgraduate College No.1, Higher Education Department, Abbottabad, Pakistan
§School of Information Technology, Deakin University, Geelong, Australia
¶Center for Artificial Intelligence, School of Computer Science, FEIT, University of Technology, Sydney, Australia
Email:*ahmedsa@ksau-hs.edu.sa,‡saeedanaz292@gmail.com §imran.razzak@ieee.org, ¶mukesh.prasad@uts.edu.au

***Abstract*—Text in natural scene image portrays rich semantic information that plays an important role in content analysis. However, apart from Arabic text in documents, the text in natural scene images exhibit much higher diversity and variability, especially in uncontrolled circumstances. In this paper, a hybrid feature extraction approach is presented to detect extremal region of Arabic scene text. The binary image and image mask are considered as a variant of input image and look for concurrent extremal regions in both images. After determination of conjoined extremal points, the scale invariant technique is applied to consider those invariant points which are common in both images based on their coordinate positions. To evaluate the performance, a multidimensional long short term memory (LSTM) network is adapted and obtained 94.21% accuracy for word recognition on unconstrained Arabic scene text recognition (ASTR) dataset.**

***Index Terms*—Extremal regions, Invariant features, Multi-dimensional LSTM, Text Recognition, Natural scene image**

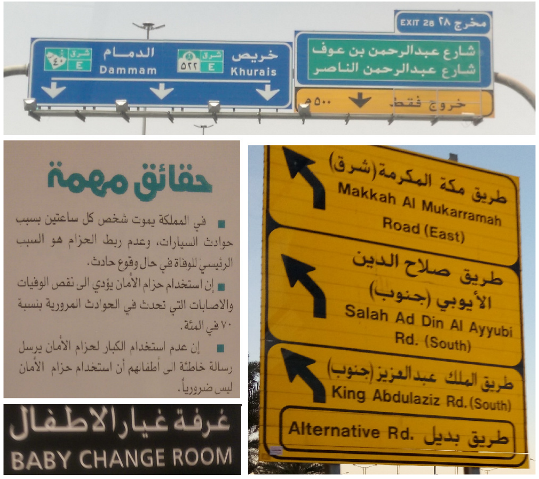

مخرج ٢٨ EXIT 28
الدمـــام
Dammam
خريص
Khurais
شارع عبدالرحمن بن عوف
شارع عبدالرحمن الناصر
خروج فقط
٥٠٠م
حقائق مهمة
في المملكة يموت شخص كل ساعتين بسبب حوادث السيارات، وعدم ربط الحزام هو السبب الرئيسي للوفاة في حال وقوع حادث.
إن استخدام حزام الأمان يؤدي الى نقص الوفيات والاصابات التي تحدث في الحوادث المرورية بنسبة ٧٠ في المئة.
إن عدم استخدام الكبار لحزام الأمان يرسل رسالة خاطئة الى أطفالهم أن استخدام حزام الأمان ليس ضرورياً.
غرفة غيار الاطفال
BABY CHANGE ROOM
طريق مكة المكرمة (شرق)
Makkah Al Mukarramah Road (East)
طريق صلاح الدين الأيوبي (جنوب)
Salah Ad Din Al Ayyubi Rd. (South)
طريق الملك عبدالعزيز (جنوب)
King Abdulaziz Rd. (South)
Alternative Rd. طريق بديل

Fig. 1: Sample Arabic scene text images

I. Introduction

The substantial demand of autonomous computing applications and it's content based image analysis have obtained popularity since recent years [1]–[3]. The detection and recognition of text in natural scene images refer to the problem of recognizing the text that appears on sign boards, posters, product packaging and billboard etc. Text in natural scene images convey rich and high-level semantic information that appear to be important for some applications to interpret. These images have significant importance for many applications such as driver assistance, number plate recognition, image based content retrieval, scene understanding and navigation aid for robotic systems etc [2], [4]–[6]. Despite great success of Optical Character Recognition (OCR) applications in most of the world languages, text recognition in natural images represented as a formidable field [7], [8] The content of an image depicts the intuitive information that may relate to variant challenges. Among various challenges the most prominent are, orientation, size, thickness, irregular shapes, different writing style and color of text in a scene image. Moreover, the text could be represented in curve shapes (i.e., characters are placed along curves rather than straight lines), and could be perspective (i.e., captured side-view).

The scene text recognition is divided into three phases including text segmentation/ localization, text extraction and recognition [1], [7], [9]. The intensive pre-processing is required to accomplish the task. In first phase, text segmentation or localization meant for detection of text area in presence of other objects in an image, while extraction means to read carefully the text and separate it in an image, so that it may later forward in a specified format to the classifier. It is obvious, that OCR system does not directly process the video images, because OCR images are usually captured in constrained scenario, it means that the clean document images capture at standard resolution and in a specific settings. However, the characteristics of scene text is different altogether, because images are captured under various conditions, thus to make it challenging for recognition systems [8]. Text images often have color blending, blur visuals, low resolution and complicated background in presence of other objects which effect the performance of any type of proposed algorithm. The techniques for scene text recognition still struggling towards accuracy because of invariant nature of text image, but some recent development proved better performance [10].

In Arabic script, there are reliance characters that make a meaningful word. The four various representation of a single character pose a challenge, thus intra-class variability may be

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

visible in many scene text images due to environmental constraints. The approaches for scene text detection/recognition tackle a much broader set of problems e.g., non-planar surfaces, unknown layout, blur, varying distance to camera, much broader resolution ranges. Various synergies exist with other methods which handle these particular impediments as a subset of elicited problems.

The **contribution** of this paper is highlighted as follows,

1) This paper presents a novel idea by concurrent invariant points approach to determine the feature points in conjoined extremal regions of Arabic word.
2) An adapted multidimensional long short term memory (MDLSTM) network is employed. The proposed methodology adapts the MDLSTM architecture in a way so that it can process invariant extremal features.
3) The database plays a vital role in evaluation of state-of-the-art techniques. There are several publicly available datasets for Latin script in the field of scene text recognition [11]–[13] but literature review suggests that there is no such dataset available for Arabic scene text. The benchmark dataset availability is a fundamental requirement for training and testing state-of-the-art classifiers. Therefore, the acquisition of text images, development of scene text database and its distribution to the researchers for comparison of different techniques and methods are main focus of attention.

The rest of the paper is organized as follows: The related work is summarized in section II. The details about proposed Arabic text detection technique and description about classifier are presented in section III. The experimental analysis and comparison of presented work with other Arabic scene text recognition are described in section IV, whereas section V concludes the proposed work.

II. Related Work

This section compiles recent development for Arabic text recognition in natural scene images. An interesting work on feature extraction is proposed by Kim et. al [1]. They categorized the text features into low-level features, high-level features and the features they obtained by merging them. The low-level features included three sub-features i.e., intensity of local variation, colors and to determine color continuity in same text area. Whereas, high-level features were used for verification of text area determined by stroke information. The performance was evaluated on variant sizes of text to determine the under considered area as text or non-text using Support Vector Machines (SVM). They have evaluated their proposed technique and reported 88.6% accuracy. Their model can be adapted and be applied on camera captured text images.

The Dynamic Time Wrapping (DTW) approach was proposed by [9] for Arabic text detection and recognition. The said approach already proved good results on huge vocabularies. One of the major ability of DTW is to learn and recognize connected and cursive characters appeared in words without explicit segmentation. It also presented the better accuracy results on character recognition in images having implicit noise. The only drawback of using their approach is computation time, which ultimately impact overall evaluation time. The Hidden Markov Model (HMM) is another choice to learn the small dataset having Arabic text. They concluded that if dataset is large, then DTW could be good choice regardless of it's time inefficiency problem. They evaluated their results on $2,000$ Arabic words and reported 97% accuracy.

The maximally stable extremal regions (MSER) is most prominent approach that was proposed for text extraction in natural images. Neumann et al [12] presents adapted MSER approach for text localization and recognition. The proposed technique was evaluated on Char74k and ICDAR2003 datasets. The accuracy reported on Char74k was 74% which is comparatively better than the accuracy they obtained on whole ICDR2003 dataset which is 57%. Toggle mapping is another technique used to separate the text from natural images which is presented by Fabrizio et al [14]. This method was initially proposed for contrast enhancement. The morphological operations were performed on gray scaled images. Their proposed solution targeted to detect the text boundary. The similar regions in an image are the focused point because this indication may help to locate the text. The proposed technique was evaluated on 501 characters which achieved 74.8% accuracy on segmented characters. Another important work on scene text detection in a wild is presented by Gomez et al [12]. They proposed hybrid approach and used MSER as text detector. They categorized their implementation into two steps, in first step they detected the text by MSER and in second step they found region of interests by using RANSAC algorithm [15]. The relation was established with text detection process and time required to inspect the regions, which concludes that time performance in module tracking increase linearly.

Another work using invariant features on extracted component is presented by Yao et al [13], they validated their process by chain analysis. They suggested component based and chain based features. On the basis of determined features like constant width, texture less and smoothing strokes, they separated text and non-text area. They evaluated their idea on proposed dataset designed for multilingual script. They divided their dataset of 500 images into 300 training and 200 test set and reported good accuracy by using their proposed technique.

The state-of-the-art technique based on deep learning Convolutional network (ConvNets) is proposed by Ahmed et al [2]. They proposed ConvNets to learn the patterns of Arabic characters. They prepared their own dataset from specialized cameras. The acquired samples were divided into trainset and testset. The 27 classes were re-scaled into 50 by 50 size. The orientation with respect to each character was also considered. In this way, they increased the size of character dataset. The 2450 samples were trained by their proposed classifier whereas, the learned network performance was reported on 250 images. They reported 0.15% character level accuracy.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

III. Arabic Text Detection and Recognition in Scene Images

Unlike Arabic script in text document, text in natural scene images exhibit much higher diversity and variability, especially in uncontrolled circumstances. The cartographic nature of Arabic script makes this task further complicated as compared to other scripts. To address the aforementioned challenges of Arabic scene text recognition in this paper, we presented a hybrid feature extraction approach to detect extremal region text image. We have considered both binary image and an image mask as a variant of input image and looked for concurrent extremal regions in both images. After determination of conjoined extremal points, the scale invariant technique is applied to consider those invariant points which are common in both images based on their coordinate positions. The proposed approach is illustrated in Figure 6.

A. Arabic Scene text (ASTR) Dataset

The dataset plays a crucial role for evaluation of state-of-the-art techniques. Its importance increases more in terms of cursive text analysis [3], [16]. It is observed from the literature [8], there is lack of efforts reported in area of Arabic scene text recognition. But some researchers are proposing Arabic scene text dataset which does not cover maximum variety of Arabic text appearances. Moreover these datasets are not publicly available [17]. By keeping this constraint forefront, this paper is also proposing Arabic Scene Text Recognition (ASTR) dataset. The dataset covers every representation of Arabic text appeared in different illumination settings with various font styles in presence of perspective distortion as represented in Figure 2. The Arabic words are segmented from acquired images. The proposed architecture and experimentation analysis are reported on the basis of segmented words.

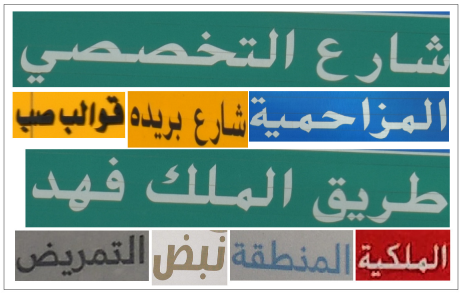
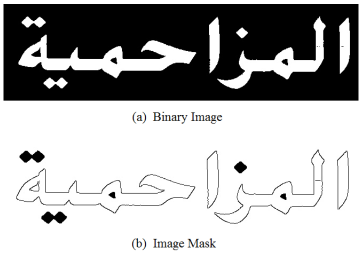


Fig. 2: Segmented Arabic textlines and words

The acquired images represent text, displayed in uncontrolled environment taken from University precinct, advertisement boards, road guides, brouchers and commodity wrappers. The captured text disintegrated into Arabic words and characters. The NikonD3300 specialized DSLR camera with 24 megapixels lens, HTC One M8 quadcore and Samsung galaxy A8 along 13 megapixels ultra pixel camera were used to capture the image having Arabic text. There are 13593 words extracted from Arabic text images. The words are further decomposed into 40597 characters. The experiments are performed at word level.

B. Invariant feature extraction from Extremal Region

This paper presents hybrid feature extraction approach that combines relevant regions of binary image and image masks by using MSER and detect invariant features by SIFT technique. All segmented text images were normalized to 50 x-height while maintaining the aspect ratio. After that, image is converted into binary and image mask as represented in Figure 3. The extremal regions were detected and keypoints were extracted from both images as indicated in the process represented in Figure 4.



(a) Binary Image

(b) Image Mask

Fig. 3: Normalized image in binary and image mask

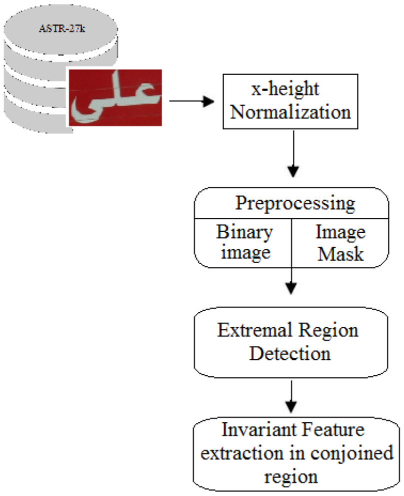

ASTR-27k
على
x-height Normalization
Preprocessing
Binary image
Image Mask
Extremal Region Detection
Invariant Feature extraction in conjoined region

Fig. 4: Pre-processing steps

The extremal region is considered as most commonly used text detection technique from natural images. It looks for minimum or a maximum values in a blob. The MSER selects that region where connected component was computed and represented uniform intensity whereas, at outside, the selected region would have contrasting background. The co-variant points were detected and merged to make a region. During the process, if points remain same even in wide range of threshold values then it would be selected as a region.

The selected region R, shows minimum invariance and affine transformation which is calculated by using the pixels involved in the process as represented in following equation,

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

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

$f(x)$ is an affine transformation function that represents a linear attitude R and a transformation variable T.

The MSER function applies threshold over the whole image, then find extremal regions through connected component. The extremal region is selected or rejected on the grounds of covering maximum or minimum area depends on whatever the criteria followed.

The invariant features were selected by considering SIFT approach which is based on Laplacian pyramid that uses various level of difference of Gaussian (DoG) function D(x,y,δ) as represented in equation 2, where δ represents Gaussian kernel of certain width and x, y show particular coordinate values.

$$D(x,y,\delta) = (G(x,y,\delta_k) - G(x,y,\delta)) * I(x,y) \quad (2)$$

where,

$$G(x,y,\delta) = \frac{1}{2\delta^2} exp[-\frac{x^2+y^2}{2\delta^2}] \quad (3)$$

By Laplacian pyramid L, as represented in equation 4, high frequency information of an image can easily be obtained because features mostly resides on these parts of an image.

$$L(x,y,\delta) = G(x,y,\delta) * I(x,y) \quad (4)$$

At each level of DoG, keypoints were detected based on the neighboring pixels in a window and select the maximum among all involved pixels in $I(x,y)$. The unnecessary keypoints that were detected on low contrast image would be rejected.

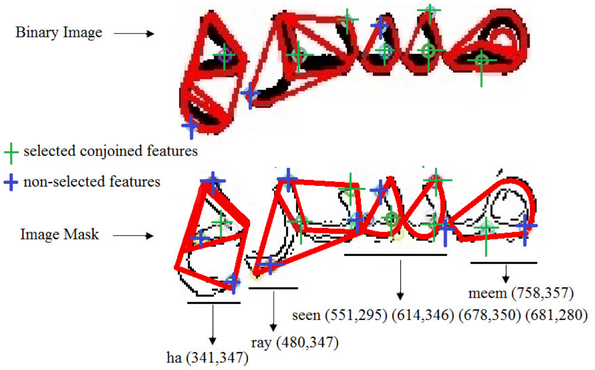


Fig. 5: Feature extraction in co-occurrence extremal regions

Figure 5 is a depiction of proposed feature selection method. Those keypoints are considered that appeared on same extremal region's location in both images i.e., binary image and image masks. The name of character and participating keypoint's coordinate values were given to classifier to be learned as a feature. The selected keypoints in co-occurrence regions in Figure 5 are represented with green plus symbol whereas, blue plus symbol indicates the non-selected features.

The important characteristic of an extremal region is the continuous affine transformation, hence this feature some times could not be able to extract exact region of interest (RoI) that requires considerations from research communities to work for further precision.

C. *Classifier for proposed architecture*

The proposed hybrid feature extraction approach was evaluated on multidimensional long short term memory (MDLSTM) networks. The MDLSTM network is used because of its strong ability to learn the sequence as it considers more appropriate in sequence learning tasks specially in cursive scripts. In recent years, MDLSTM proved better performance on cursive scripts as reported in [2], [3], [18]. The LSTM deals with memory blocks reside in hidden layer instead of memory units. The history maintains with the help of multiplicative units exists in memory blocks.

As represented in Figure 6, the labels are provided with corresponding feature coordinates to MDLSTM classifier. To get better performance, the number of hidden layers, learning rate and momentum values are empirically selected. The Connectionist Temporal Classification (CTC) process the learned sequence generated by neural network classifier for a purpose to label the unsegmented data. The CTC is independent layer attached to classifier's output so that it may predict the learned output based on scoring function. It is usually used to predict the sequences of unsegmented data (like in cursive script) that has been trained by recurrent neural network. The CTC output mapped through softmax layer which take the training output into account to model the prediction of labels with corresponding coordinate values in particular. The input of CTC are sequences that have been observed during training and output are the probabilities of sequential labels. The difference of observation score in CTC is back propagated to update the weights of neural network.

IV. Results

The conducted experiments were performed on word-level recognition. The model is trained on Arabic words and measure the performance of network on Arabic word dataset. As implicit segmentation is performed by MDLSTM networks, so that unsegmented data is presented to the classifier for learning. The training model learns the label of a provided sample with its coordinate values which appeared in co-occurrence regions. The coordinate values which are not appeared on similar location in both images will be disregarded as a feature.

TABLE I: Parameter's Statistics

Parameter's detail	Values
Number of memory blocks in each hidden layer	20,40,60,80,100
Hidden block Size	2×5
Input block size	1×20
Learning rate	1×10^{-3} / 1×10^{-2}
Momentum	0.7/0.9

The experiments were performed with number of different hidden layer sizes, learning rate, and momentum value. Table I displayed the final parameter's values that have been selected empirically. These parameters determined the best reported accuracy.

Figure 7, depicts the learning curve observed during training process. The experiments were performed with different learning rates and hidden layer memory blocks. The best training

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

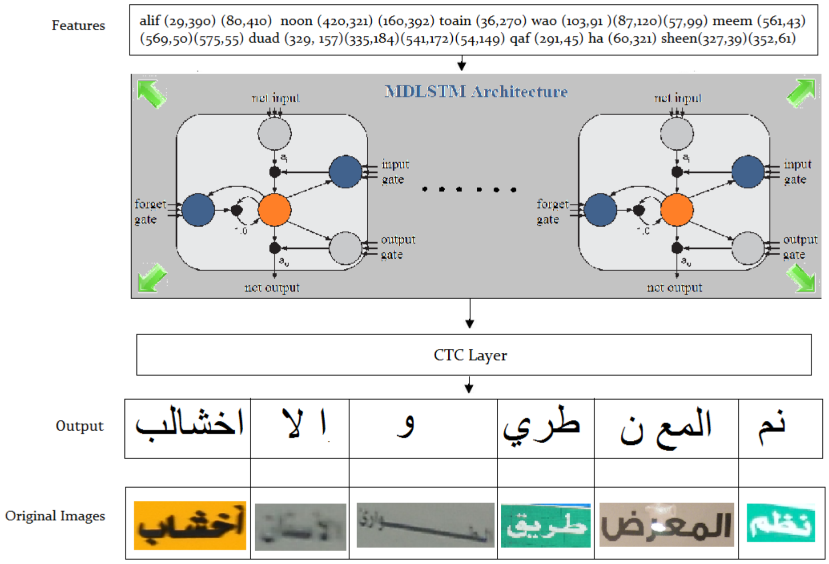

Features
alif (29,390) (80,410) noon (420,321) (160,392) toain (36,270) wao (103,91)(87,120)(57,99) meem (561,43)
(569,50)(575,55) duad (329, 157)(335,184)(541,172)(54,149) qaf (291,45) ha (60,321) sheen(327,39)(352,61)
MDLSTM Architecture
net input
input gate
forget gate
output gate
net output
CTC Layer
Output
اخشالب
ا لا
و
طري
المع ن
نم
Original Images

Fig. 6: Proposed methodology and expected output depiction for good or bad images

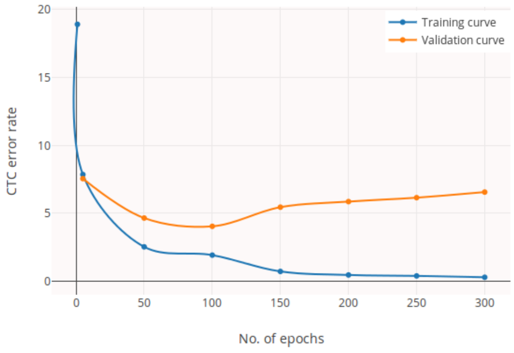


Fig. 7: Best Learning network curve observed on 100 LSTM memory blocks

accuracy was obtained when hidden layer had 100 memory blocks. As observed from learning curve initially, the validation curve dropped down but after 100 iterations the curve increased. When the difference between training and validation curve increased the training was stopped. Approximately 96% learning accuracy is obtained during training.

A. *Discussion*

In this section, the comparison of presented work is conducted with recently proposed work on Arabic text analysis either appears online, offline or as a scene text images. The reported accuracy in presented work is 94% on Arabic scene word recognition.

The work presented by Abandah etal [19] is one of the important work to recognize Arabic handwritten words. They proposed efficient segmentation approach that achieved the high accuracy using recurrent neural network. They come to the conclusion that their segmentation approach with efficient feature extraction provides good result in comparison to holistic approach that extracts features from raw pixels. They improved the results presented on Arabic word recognition in ICDAR 2009 using MDLSTM. They evaluated their work on JU-OCR2 dataset with 98.96% accuracy on word level. They also assessed the potential of their proposed technique with IFN/ENIT dataset of Arabic handwritten words. Another interesting work on handwritten Arabic recognition systems is presented by Ahmad et al [20]. They proposed feature extraction technique with adapted sliding window approach that consider the handwritten Arabic text by ink-pixel distribution. Their technique is applicable for multifont styles. They performed experiments on two separate dataset to check the performance of their proposed technique with the combination of Hidden Markov Models (HMM). They evaluated their proposed architecture on APTI [21] dataset and achieved 96.99% accuracy.

Another work on the Arabic character recognition using deep learning ConvNets approach is presented by Ahmed et al [2]. They consider each shape of 27 classes exists in Arabic with five orientations. They formulated their experiments by considering the filter size 3 and 5 with stride value 1 and 2. They reported 85% accuracy on character level. Jain et al [22] compiled the research work on Arabic video text images. They considered RNN approach as it is best suited for sequential learning problems like we have in cursive Arabic script. They evaluated their proposed method by hybrid approach that consist of ConvNets and RNN. The experiments were performed on ALIF dataset which have Arabic video text samples and their own dataset captured from camera. Their own collected data samples integrated into words and then characters. They reported character level and word level accuracy on video rendered text and a segmented text from captured image. They achieved 98.17% accuracy on character and 79.67% accuracy on word recognition in ALIF dataset.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE II: Performance comparison of Arabic dataset analysis

Method	Arabic dataset type	Dataset	Accuracy (%)
G. Abandah et al [19]	Offline Handwriting	JU-OCR2	98.60
I. Ahmed et al [20]	Printed text	APTI	96.99
SB. Ahmed et al [2] [14]	scene text	ASTR	85.0
M. Jain et al [22]	video text	ALIF	98.17 on characters/ 79.67 on words
M. Jain et al [22]	scene text	2000-words	75.05 on characters/ 39.43 on words
The Proposed Method	scene text	ASTR-28k	94.01

Hence, on camera captured images they reported 75.05% and 39.43% accuracy on character and word level respectively.

V. Conclusion

This paper presents the idea of concurrent invariant features points that are detected on cursive scene text images, appeared in detected extremal regions. The novel method is demonstrated to quantifying the co-occurrence of invariant features, reside in extremal regions. The performance has been measured on Arabic words that were segmented from natural images. The evaluation results shows benchmark accuracy as far as Arabic scene text recognition is concerned. As most of the recent work on Arabic scene text focused on video rendered text so in that context, the presented work is a benchmark effort in Arabic scene text analysis. The comparison is performed with other reported state-of-the-art techniques and established the fact that the achieved reported accuracy is the best one even compared with other cursive scripts. The error rate of 0.6% is reported on Arabic word recognition in natural images. Moreover in this paper, the constraints of Arabic and Arabic like scripts are highlighted, so that to find the possibilities of future research in the field of Arabic scene text. Another possible exploitation of presented idea might be to learn the features and transfer the most relevant features to subsequent layers.

References

[1] K. C. Kim, H. R. Byun, Y. J. Song, Y. W. Choi, S. Y. Chi, K. K. Kim, and Y. K. Chung, "Scene text extraction in natural scene images using hierarchical feature combining and verification," in *ICPR*, 2004, pp. II: 679–682. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.1334350

[2] S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, "Deep learning based isolated arabic scene character recognition," in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. IEEE, 2017, pp. 46–51.

[3] S. B. Ahmed, S. Naz, S. Swati, and M. I. Razzak, "Handwritten urdu character recognition using one-dimensional blstm classifier," *Neural Computing and Applications*, vol. 31, no. 4, pp. 1143–1151, 2019.

[4] Y. Qian and Z. Li, "A method of multi-license plate location in road bayonet image," *International Journal of Advanced Research in Artificial Intelligence(IJARAI)*, vol. 4, no. 4, 2015. [Online]. Available: http://dx.doi.org/10.14569/IJARAI.2015.040404

[5] T. Wilhelm, "Towards facial expression analysis in a driver assistance system." IEEE, 2019, pp. 1–4.

[6] S. B. Ahmed, M. I. Razzak, and R. Yusof, *Cursive Script Text Recognition in Natural Scene Images: Arabic Text Complexities*. Springer Nature, 2019.

[7] A. A. Shahin, "Printed arabic text recognition using linear and nonlinear regression," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 8, no. 1, 2017.

[8] S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yusof, "Arabic cursive text recognition from natural scene images," *Applied Sciences*, vol. 9, no. 2, 2019. [Online]. Available: http://www.mdpi.com/2076-3417/9/2/236

[9] M. Khemakhem and A. Belghith, "Towards a distributed arabic ocr based on the dtw algorithm: performance analysis," *Int. Arab J. Inf. Technol*, vol. 6, no. 2, pp. 153–161, 2009.

[10] S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yusof, "Cursive scene text analysis by deep convolutional linear pyramids," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 307–318.

[11] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *ACCV (3)*, ser. Lecture Notes in Computer Science, R. Kimmel, R. Klette, and A. Sugimoto, Eds., vol. 6494. Springer, 2010, pp. 770–783.

[12] L. Gomez and D. Karatzas, "Mser-based real-time text detection and tracking," in *ICPR*. IEEE Computer Society, 2014, pp. 3110–3115. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6966883

[13] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *CVPR*. IEEE Computer Society, 2012, pp. 1083–1090. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6235193; http://www.computer.org/csdl/proceedings/cvpr/2012/1226/00/index.html

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

[15] J. Matas and O. Chum, "Randomized RANSAC with td,d test," *Image and Vision Computing*, vol. 22, no. 10, pp. 837–842, Sep. 2004. [Online]. Available: http://www.sciencedirect.com/science/article/B6V09-4CP6H4K-1/2/0239fa040b97225f7c25f2c10e51135a

[16] S. Naz, A. I. Umar, R. Ahmad, S. B. Ahmed, S. H. Shirazi, and M. I. Razzak, "Urdu nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features," *Neural Computing and Applications*, vol. 28, no. 2, pp. 219–231, 2017.

[17] S. Yousfi, S. Berrani, and C. Garcia, "Alif: A dataset for arabic embedded text recognition in tv broadcast," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1221–1225.

[18] S. Naz, A. I. Umar, R. Ahmad, I. Siddiqi, S. B. Ahmed, M. I. Razzak, and F. Shafait, "Urdu nastaliq recognition using convolutional-recursive deep learning," *Neurocomputing*, vol. 243, pp. 80–87, 2017.

[19] G. A. Abandah, F. T. Jamour, and E. A. A. Qaralleh, "Recognizing handwritten arabic words using grapheme segmentation and recurrent neural networks," *IJDAR*, vol. 17, no. 3, pp. 275–291, 2014.

[20] I. Ahmad, S. A. Mahmoud, and G. A. Fink, "Open-vocabulary recognition of machine-printed arabic text using hidden markov models," *Pattern Recognition*, vol. 51, pp. 97–111, 2016.

[21] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert, "A new arabic printed text image database and evaluation protocols," in *ICDAR*, 2009, pp. 946–950. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2009.155

[22] M. Jain, M. Mathew, and C. V. Jawahar, "Unconstrained scene text and video text recognition for arabic script," *CoRR*, vol. abs/1711.02396, 2017. [Online]. Available: http://arxiv.org/abs/1711.02396

978-1-7281-6926-2/20/$31.00 ©2020 IEEE