ICPR2020 Competition on Text Detection and Recognition in Arabic News Video Frames

Oussama Zayene[1[0000−0001−9529−925X]], Rolf Ingold[2[0000−0001−7738−133X]], Najoua Essoukri BenAmara[3[0000−0001−7914−0644]], and Jean Hennebert[1[0000−0002−5616−6830]]

[1] Institute of Complex Systems, HES-SO//Fribourg, Switzerland
{oussama.zayene, jean.hennebert}@hefr.ch
[2] DIVA group, Department of Informatics, University of Fribourg, Switzerland
rolf.ingold@unifr.ch
[3] LATIS lab., National Engineering School of Sousse, University of Sousse, Tunisia
najoua.benamara@eniso.rnu.tn

Abstract. After the success of the two first editions of the "Arabic Text in Videos Competition—AcTiVComp" , we are proposing to organize a new edition in conjunction with the 25th International Conference on Pattern Recognition (ICPR'20). The main objective is to contribute in the research field of text detection and recognition in multimedia documents, with a focus on Arabic text in video frames. The former editions were held in the framework of ICPR'16 and ICDAR'17 conferences. The obtained results on the AcTiV dataset have shown that there is still room for improvement in both text detection and recognition tasks. Four groups with five systems are participating to this edition of AcTiVComp (three for the detection task and two for the recognition task). All the submitted systems have followed a CRNN-based architecture, which is now the de facto choice for text detection and OCR problems. The achieved results are very interesting, showing a significant improvement from the state-of-the-art performances on this field of research.

Keywords: Text Detection · Text Recognition · Arabic News indexing · AcTiVComp · ICPR competition.

1 Introduction

Among the pattern recognition fields, automatic text recognition, known as OCR, has been widely studied for its prominent position in our everyday life. OCR has a long history of research that started from isolated character recognition and evolved to printed/handwriting document recognition. Over the last decade, embedded texts in videos and natural scenes have received increasing attention as they often give crucial information about the media content [9]. Nevertheless, extracting text from such content is a non-trivial task due to many challenges like low resolution, background complexity and text variability in terms of size, color and font. All these challenges may give rise to failures in video text detection and recognition tasks.

Over the last decade, interest in this area of research has led to a plethora of text detection and recognition methods. So far, these methods have focused only on Latin and Chinese characters. For a language like Arabic, which is used by more than one billion people around the world, the literature concerning video text analysis is limited to few studies [2, 5, 10]. In contrast to non Arabic text where most of the methods have been tested and compared in the context of international competitions, e.g., ICDAR Robust Reading Competition (RRC) [4], most of the existing methods for Arabic video text detection/recognition were tested on private datasets with non-uniform evaluation protocols. This makes direct comparison and scientific benchmarking rather impractical.

The present competition[5] aims to fill the aforementioned gap by encouraging Arabic Video OCR researchers to develop and test their systems on a standard dataset and using the same evaluation metrics.

This contest represents a part of the AcTiVComp series that have been organized respectively within the ICPR'16 [19] and ICDAR'17 [20] conferences. Actually, the former editions have attracted seven groups for participating and have received ten systems in total. The best achieved F-score for the channel-free detection protocol was 0.85 (more details about the protocols are explained in Section 3). For the recognition task, the best results in the channel-free protocol have not exceeded 0.76 in terms of Line Recognition Rate (LRR). Furthermore, the obtained results on the recently added subset (SD 480x360) were quite low for almost all the participating systems. For this reason, we are organizing a new edition so as to improve these results, especially for the channel-free evaluation protocols.

AcTiVComp has been organized by iCoSys Institute[6] from the University of Applied Sciences and Arts, Western Switzerland and LATIS Lab[7] from the National Engineering School of Sousse, Tunisia, in collaboration with DIVA Group[8], from the University of Fribourg, Switzerland.

The participants to this third edition of AcTiVComp had roughly five months to train their systems before the test data was released. After two additional weeks the teams had to submit their results.

In the following, we first present the used datasets in section 2. Section 3 is dedicated to the competition tasks. We describe the participating systems in section 4. The results are discussed in section 5 and section 6 draws the conclusions.

2 Competition Datasets

AcTiV is a real-content database where video clips are recorded from four Arabic news channels, TunisiaNat1, France24 Arabic, Russia Today Arabic and Al-

[4] https://rrc.cvc.uab.es/
[5] https://diuf.unifr.ch/main/diva/AcTiVComp/
[6] https://icosys.ch
[7] http://www.latis-eniso.org
[8] https://www3.unifr.ch/inf/diva/en/

Fig. 1: Typical video frames from AcTiV-D dataset. From left to right : Examples of RussiaToday Arabic, France24 Arabe, TunisiaNat1 and AljazeeraHD

Table 1: Detection Dataset and Evaluation Protocols

Protocol	**TV Channel**	**Training set**	**Test set**	**Closed-test set**
		#Frames	#Frames	#Frames
1	AljazeeraHD	337	87	103
4	France24 arabe	331	80	104
	RussiaToday arabic	323	79	100
	TunisiaNat1	492	116	106
	All SD channels	1,146	275	310
4bis	TunisiaNat YouTube	-	150	149
7	All channels	1,483	362	413

jazeeraHD, using a DBS system and then transcoded and segmented into frames. It was presented in the ICDAR 2015 conference [21] and then in the Journal of Imaging 2018 [23] as the first publicly accessible annotated dataset designed to assess the performance of different Arabic Video OCR systems. The two main challenges addressed by this dataset are the following:

- The variability of text patterns, e.g. text colors, fonts, sizes and position.
- the presence of complex backgrounds with various text-like objects. It is currently used by more than 40 research groups around the world.

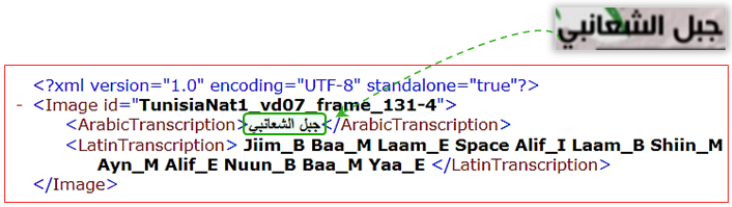
AcTiV includes two appropriate datasets, namely AcTiV-D and AcTiV-R, for detection and recognition tasks. These datasets were used as a benchmark in the two first editions of AcTiVComp.

2.1 AcTiV-D

AcTiV-D represents a dataset of non-redundant frames used to build and evaluate text detection methods. It contains a total of 2,557 news video frames that have been hand-selected with a particular attention to achieve a high diversity in text regions. Figure 1 states some examples from AcTiV-D for typical problems in video text detection. ActiV-D frames are distributed over four sets (one set per TV channel). Every set includes three sub-sets: training set, test set and closed-test set (used for competitions only). The statistics are presented in Table 1. ActiV-D includes some frames that do not contain any text and some others

```
<?xml version="1.0" encoding="UTF-8"?>
- <Protocol4 channel="TunisiaNat1">
  - <frame id="7" source="vd01">
      <rectangle id="1" x="506" y="464" width="61" height="14"/>
      <rectangle id="2" x="66" y="499" width="491" height="32"/>
    </frame>
  - <frame id="16" source="vd01">
      <rectangle id="1" x="441" y="464" width="127" height="18"/>
      <rectangle id="2" x="373" y="499" width="184" height="27"/>
    </frame>
  - <frame id="64" source="vd01">
      <rectangle id="1" x="429" y="462" width="138" height="24"/>
    </frame>
```

(a)



(b)

Fig. 2: (a) Part of detection XML file of TunisiaNat1 TV channel. (b) Recognition ground-truth file and its corresponding textline image.

Fig. 3: Example of text images from AcTiV-R dataset depicting typical characteristics of video text images

that contain the same text regions but with different backgrounds. The detection ground-truth is provided at the line level for each frame. Figure 2a depicts a part of the ground-truth XML file of protocol4.3 (TunisiaNat TV channel). One bounding box is described by the element *rectangle*, which contains the rectangle attributes: (x, y) upper-left coordinates, width and height.

2.2 AcTiV-R

AcTiV-R is a dataset of textline images used to build and evaluate Arabic text recognition systems. Different fonts, sizes, backgrounds and colors are represented in the dataset. Figure 3 illustrates typical examples from AcTiV-R. The collected text images cover a broad range of characteristics that distinguish video

Table 2: Recognition Dataset and Evaluation Protocols. *Lns*, *Wds*, *Chars* and *YT* respectively denote *Lines*, *Words*, *Characters* and *YouTube*.

Protocol	**TV Channel**	**training set**			**test set**			**closed-test set**		
		Lns	Wds	Chars	Lns	Wds	Chars	Lns	Wds	Chars
3	AlJazeeraHD	1,909	8,110	46,563	196	766	4,343	262	1,082	6,283
6	France24 arabe	1,906	5,683	32,085	179	667	3,835	191	734	4,600
	RussiaToday	2,127	13,462	78,936	250	1,483	8,749	256	1,598	9,305
	TunisiaNat1	2,001	9,338	54,809	189	706	4,087	221	954	5,597
	All SD channels	6,034	28,483	165,830	618	2,856	16,671	668	3,286	19,502
6bis	TunisiaNat1 YT	-	-	-	320	1,487	8,726	311	1,148	6,645
9	All channels	7,943	36,593	212,393	814	3,622	21,014	930	4,368	25,785

frames from scanned documents. AcTiV-R consists of 10,415 textline images, 44,583 words and 259,192 characters. The recognition ground-truth is provided at the line level for each text image. Figure 2b depicts an example of a ground-truth XML file and its corresponding textline image. The file is composed of two principal markup sections: ArabicTranscription and LatinTranscription. To have an easily accessible representation of Arabic text, it is transformed into a set of Latin labels with a suffix that refers to the letter's position in the word, i.e. B: Begin, M: Middle, E: End and I: Isolate. During the annotation process, 164 character shapes were considered, including 10 digits and 12 punctuation marks. The statistics are presented in Table 2.

3 Competition Tasks

AcTiVComp20 includes three main tasks: i) Text detection, ii) text recognition and iii) end-to-end text recognition in Arabic news video frames. Each of these tasks may include one or more evaluation protocols. Only the two first tasks are described in what follows, as the third one has not received any submission. Please refer to the competition website[9] for more details about this task.

3.1 Task 1: Text Detection

The objective of this task is to obtain an estimation of text regions in a video frame in terms of bounding boxes (x, y, width and height). In the following paragraphs we present details about the used evaluation protocols and metrics for this task.

Evaluation Protocols:

– **Protocol 1** aims to measure the performance of text detection methods in HD frames.

[9] `https://diuf.unifr.ch/main/diva/AcTiVComp/`

– **Protocol 4** is similar to protocol 1, varying only in channel resolution. All SD (720x576) channels in AcTiV dataset are targeted by this protocol, which is split in four sub-protocols: three *channel-dependent* protocols (p4.1, p4.2 and p4.3) and one *channel-free* protocol (p4.4).

– **Protocol 4bis** is dedicated to the last added resolution (480 x 360). The main idea of this protocol is to train a given system with SD (720 x 576) data, i.e. Protocol 4.3, and test it with different data resolution and quality.

– **Protocol 7** is the generic version of protocols 1 and 4 where text detection is evaluated regardless of data quality.

Table 1 summarizes the detection protocols.

Metrics: Following the evaluation metrics of AcTiVComp's previous editions and those of ICDAR RRC series, the text detection task of AcTiVComp20 is evaluated in terms of Precision, Recall and F-measure that are calculated as

$$Precision = \frac{\sum_i match_D(D_i, G, t_r, t_p)}{|D|} \tag{1}$$

$$Recall = \frac{\sum_j match_G(G_j, D, t_r, t_p)}{|G|} \tag{2}$$

$$Fmeasure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

where D is the list of detected rectangles, G is the list of ground-truth rectangles, $match_D$ and $match_G$ are the matching functions, and t_r and t_p are two quality constraints on area recall and area precision respectively. In the experiments, t_r is fixed to 0.8 and t_p is fixed to 0.4 [10]. These measures are calculated using the evaluation tool presented in [22], which takes into account all types of matching cases between G and D bounding boxes, i.e. one-to-one, one-to-many and many-to-one matching.

3.2 Task 2: Text Recognition

Taking a textline image as input, the objective of this task is to generate the corresponding text transcriptions. The used evaluation protocols and metrics are presented below.

[10] This choice is motivated by the fact that a detection result which cuts parts of the text rectangle is more disturbing than a detection which results in a too large rectangle.

Evaluation Protocols:

- **Protocol 3** is dedicated to evaluate the performance of OCR systems to recognize text in HD frames.
- **Protocol 6** is similar to protocol 3, differing only in the channel resolution. All SD (720x576) channels in AcTiV dataset are targeted by this protocol, which is split in four sub-protocols: three *channel-dependent* protocols (p6.1, p6.2 and p6.3) and one *channel-free* protocol (p6.4).
- **Protocol 6bis** is dedicated to last added resolution (480x360) for TunisiaNat1 TV. The idea behind is to train a given system with SD (720x576) data and test it with different data resolution and quality.
- **Protocol 9** is the generic version of the previous protocols where text recognition is assessed without considering data quality.

Table 2 presents these protocols in more details.

Metrics: The performance measure for the recognition task is based on the Line Recognition Rate (LRR), and on the computation of Insertion (I), Deletion (Dl) and Substitution (S) errors at the level of Character Recognition Rate (CRR). CRR and LRR are calculated as

$$CRR = \frac{\#characters - I - S - Dl}{\#characters} \tag{4}$$

$$LRR = \frac{\#lines\_correctly\_recognized}{\#lines} \tag{5}$$

4 Submitted Methods

Overall, 5 methods from 4 different teams were submitted for the two first tasks of AcTiVComp challenge. All the methods followed a CNN-based architecture, which is now the de facto choice for text detection and recognition problems.

4.1 Text-Fcos

The Text-Fcos method is submitted for the detection task by Michael Jungo, Beat Wolf and Andreas Fischer from the School of Engineering and Architecture of Fribourg (HEIA-Fr), Switzerland.

The used model is based on FCOS [15], a one-stage anchor-free object detector. EfficientNet [14] is used as a backbone with two key changes: i) Group Normalization [17] instead of Batch Normalization. ii) Every point-wise convolution (i.e. 1×1 convolution) has been replaced by Ghost Convolutions [3] with $ghost\_factor = 2$, where $\frac{1}{2}$ of the features are generated by a 1×1 convolution

and the remaining $\frac{1}{2}$ are created from the resulting features by applying a 3×3 depth-wise convolution to them.

For the training phase, the Adaptive Training Sample Selection (ATSS) [24] has been employed with $k = 9$, which selects the 9 best candidates per Feature Pyramid Network (FPN) level for each ground-truth bounding box. Regarding the loss functions, the Focal Loss [8] was used for the classification, the Generalized Intersection over Union Loss (GIoU) [12] was used for the bounding box regression and the binary cross-entropy loss for the centerness. Only the provided AcTiV-D train dataset has been used for training, but for each image, 10 additional images have been generated by applying random augmentations to that image, resulting in 13,768 images. All images have been resized such that the larger side is 768 pixels while preserving the aspect ratio, regardless of the image resolution. The training was performed on two Titan RTX with mixed-precision. For the B4 model (EfficientNet-B4 as backbone), a batch size of 12 per GPU was used and one epoch took roughly 12m30s. It was trained for a little over 70 epochs, with a total of 15h.

During inference, only bounding boxes with a classification probability over 5% are considered, but this can result int false-positives. In order to alleviate this problem, a dynamic threshold is calculated based on the mean and standard deviation of the possible bounding box locations. This threshold removes a lot of low quality bounding boxes, which many times do not contain any text, but that also removes some of the bounding boxes containing actual text, simply having a low confidence.

4.2 EffDB-UNet

The EffDB-UNet text detection system is submitted for the detection task by Lokesh Nandanwar and Shivakumara Palaiahnakote which are members of Multimedia Lab, Faculty of Computer Science and Information Technology at the Univerity of Malaya, Malaysia; Ramachandra Raghavendra from NTNU, Norway and Umapada Pal from CVPR Unit, ISI Kolkata, India.

The submitted system contains mainly two stages, namely Deep CNN model and post-processing step. For the first stage, the normalized input frame is passed through the EffDB-UNet model. This model is based on the combination of three major components: EfficientNet Backbone (B4) [14] as Encoder, UNet as Decoder [13] and differentiable binarization (DB) as a head. Inspired by [7], the adaptive thresholding and the DB of output mask are then applied to get the desired output. The EffDB-UNet model takes 3 channel input and gives 2 channel output consisting of segmentation mask and border mask of the same size as the input. In the second stage, the label is generated for the outputs, inspired by the Progressive Scale Expansion Network (PSENet) [16], the threshold segmentation and border mask are used to generate quadrilateral of text regions described by a set of segments with a threshold of 0.6. The model is completely trained on SynthText dataset and ICDAR19-MLT Scene Text dataset [11], and finally finetuned on the competition training dataset along with non-text data collected from ICDAR2015 scene text dataset [6]. While training the model augmentation

techniques such as East Random Cropping, Random Flipping, and Random Rotation are used.

4.3 THDL-Det

The THDL-Det system is submitted for the detection task by Shanyu Xiao, Ruijie Yan, Gang Yao, Haodong Shi and Liangrui Peng from the Department of Electronic Engineering, Tsinghua University, Beijing, China.

The system is an end-to-end text spotter based on the Mask R-CNN instance segmentation framework [4], and the text detection process can be divided into two stages. At the first stage, a CNN extracts high-level feature maps from an input image, and the region proposal network (RPN) classifies positive/negative anchors and makes regression to achieve precise location. Guided by the classification and regression objective functions in the training process, the RPN generates a set of rectangle proposal boxes for text regions. At the second stage, a varying-size RoIAlign layer is proposed to extract features for region proposals with different aspect ratios. Then two fully-connected sub-networks filter non-text regions and make more precise location predictions. A fully-convolutional network is used to predict an instance mask for each text region, and a smallest enclosing quadrilateral is constructed from the mask. The hyper-parameters in Mask R-CNN, including anchor aspect ratios, different schemes and parameters of non-maximum suppression are fine-tuned. The ResNeXt-101 [18] is used as the backbone, and the multiscale training strategy is adopted. The system is pre-trained on the SynthText dataset and the ICDAR 2019 MLT dataset [11], and fine-tuned on the AcTiVComp20 training set of text detection. The system is implemented using the PyTorch framework. Detection results of different TV channels in the AcTiVComp20 dataset are generated by a single model.

4.4 THDL-Rec

The THDL-Rec system is submitted for the recognition task by Ruijie Yan, Shanyu Xiao, Gang Yao and Liangrui Peng from the Department of Electronic Engineering, Tsinghua University, Beijing, China.

The system adopts a CNN-LSTM-CTC framework to recognize Arabic text lines in videos. For feature extraction, a modified EfficientNet-B5 [14] with a U-shaped structure is used. The original EfficientNet-B5 has seven convolutional blocks. To construct the U-shaped structure, the output of the seventh convolutional block is up-sampled and summed up with the output of the fifth convolutional block, and further up-sampled and summed up with the output of the third convolutional block. Feature maps output by the U-shaped CNN has a size of $512 \times 8 \times w$, where 512 is the number of channels and 8 is the height of the feature maps. w is the width of the feature maps, which is proportional to the width of the input image. Five additional convolutional layers and a max-pooling layer are then used to transform feature maps into a feature sequence with size $w \times 512$. Finally, the feature sequence is processed by a two-layer bidirectional LSTM network followed by a CTC decoding layer for

text transcription. The system is pre-trained on about 3 million synthetic text line images and fine-tuned on the AcTiVComp20 training set of text recognition. The synthetic text line images were generated by using the ANT Corpus [1] as text contents. The system is implemented by using the PyTorch framework on a single NVIDIA Tesla V100 GPU. Recognition results of different TV channels in the AcTiVComp20 dataset are generated by a single model. By evaluating on the whole competition test set with batch size = 1 and beam size = 5, the average recognition time on a single image is 12,8ms.

4.5 ArabOCR

The ArabOCR system is submitted for the recognition task by Abdul Rehman from the School of Electrical Engineering and Computer Science, NUST, Islamabad, Pakistan, Adnan Ul-Hasan and Faisal Shafait from the Deep Learning Laboratory, National Center of Artificial Intelligence (NCAI), Islamabad, Pakistan.

This system is based on a CRNN architecture, which consists of three parts: i) First a CNN block is used to extract features from the input text image. Each convolution layer of this block is activated with a Leaky Rectifier Linear Units (LeakyReLU) layer. Batch Normalization [29] is also used in all convolutional layers to normalize the inputs of non-linear activation functions. ii) After the CNN stage, 2 Bi-directional GRU recurrent layers (with 256 cells per layer) are applied. BatchNorm and LeakyReLU are again used here. iii) The last part contains a single convolution layer with kernel size of 1, followed by a LogSoftmax layer. Finally the CTC layer is used to decode the output.

The original images are converted into grayscale and resized to have a fixed height of 64 while keeping the aspect ratio. During training, input images have been randomly augmented. A total of 8 augmentations were (3 shape-based and 5 color-based) applied per image. The system was implemented using Pytorch Framework. The training was performed on a single Nvidia GeForce GTX1080 Ti GPU. The model took approximately 8 hours to be trained.

5 Results and Analysis

This section presents results of the submitted methods under each task along with their analysis. Final results at the end of the competition period are provided in Table 3 and Table 4. All participants in Task 1 have employed semantic segmentation methods to accurately localize text instances. THDL-Det adopted a two-stage anchor-based strategy following the Mask R-CNN framework, and used ResNeXt as backbone. EffDB-UNet adopted a two-stage anchor-free strategy that combines EfficientNet with UNet followed by PSENet as a refinement step. While Text-Fcos was built on one-stage anchor-free detector (FCOS). EfficientNet was also used here as a backbone.

The THDL-Det team achieves the best score in F-measure, precision and recall for almost all protocols. The system provides an effective F-measure of

Table 3: Results of Task 1 (Text detection). R, P and F respectively denote Recall, Precision and F-measure.

Protocol/ System		**P1**	**P4.1**	**P4.2**	**P4.3**	**P4.3bis**	**P4.4**	**P7**
Text-Fcos	R	0.83	0.88	0.91	0.91	**0.87 (1)**	0.89	0.88
	P	0.85	0.87	0.91	0.91	0.87	0.89	0.88
	F	0.84 (3)	0.88 (2)	0.91 (2)	0.91 (2)	**0.87**	0.89 (2)	0.88 (2)
THDL-Det	R	0.92	0.90	0.92	0.92	0.79	0.91	0.91
	P	0.90	0.90	0.92	0.92	0.79	0.91	0.91
	F	**0.91 (1)**	**0.90 (1)**	**0.92 (1)**	**0.92 (1)**	0.79	**0.91 (1)**	**0.91 (1)**
EffDB-UNet	R	0.89	0.83	0.76	0.77	0.86	0.79	0.79
	P	0.89	0.83	0.76	0.77	**0.88 (1)**	0.79	0.79
	F	0.89 (2)	0.83 (3)	0.76 (3)	0.77 (3)	**0.87**	0.79 (3)	0.79 (3)

Table 4: Results of Task 2 (Text recognition)

Protocol/ System		**P3**	**P6.1**	**P6.2**	**P6.3**	**P6.3bis**	**P6.4**	**P9**
THDL-Rec (1)	CRR	99.83	99.34	99.48	99.43	-	99.43	99.53
	LRR	**95.80**	**87.43**	**85.94**	**85.07**	-	**85.63**	**88.71**
ArabOCR (2)	CRR	99.49	98.31	98.72	99.07	-	98.75	98.94
	LRR	90.84	72.77	71.48	77.83	-	74.10	79.03

0.91 for the global (channel-free) protocol 7, which implies its generalization ability in detecting text regions regardless the resolution. Yet, this score has decreased by 11% in the protocol 4.3bis (SD 480x360, YouTube quality). This can be explained by the fact that such object detectors i.e., Mask R-CNN, rely heavily on predefined anchors, which are sensitive to hyper-parameters (e.g., input size, aspect ratio, scales). The Text-Fcos team takes the second place with a small difference from the winner in terms of F-measure (ranging from 1 to 3%) for all protocols except two: Protocol p1 where THDL-Det is ahead by 7% and protocol p4.3bis where Text-Fcos is ahead by 8%. EffDB-UNet team achieves good results for all protocols and gets the first place in protocol p4.3bis with Text-Fcos, and outperformed him by 5% in protocol p1.

The best result of the recognition challenge is marked in bold in Table 4. THDL-Rec system shows a superiority in all the evaluation protocols with a gain ranging from 5% to 14% compared to the ArabOCR system. It is worth to note that both systems have used a CRNN architecture in a different manner.

The best achieved results on the global protocol 9, which are around 88% in terms of line recognition rate, represent a significant improvement in the Arabic Video OCR field.

Yet, working on text detection and text recognition separately is considered less challenging than working on the end-to-end recognition where all textlines in a given input frame should be correctly localized and recognized in a single step.

We are hoping to receive more submissions in the next edition of AcTiVComp, especially for the end-to-end task.

6 Conclusions

The third edition of AcTiVComp has attracted four teams for participating in the two tasks of text detection and text recognition. As seen in the results and analysis section, the rates of the winning systems, from the THDL team, are very interesting, showing a significant improvement from the state-of-the-art performances on this research problem [20], e.g., compared to the highest rates of the previous editions of AcTiVComp, the new achieved detection F-score has increased by 6% on the global protocol p7, and for the recognition task, the new results are higher with gains of respectively 13% and 14% on the global protocols p6.4 (SD TV channels) and p9 (All TV channels). The obtained results can be further improved. Hence, we look forward to have more participants in the future editions of AcTiVComp and more researchers joining the Arabic video text detection and recognition research topic.

References

1. Chouigui, A., Khiroun, O.B., Elayeb, B.: Ant corpus: an arabic news text collection for textual classification. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). pp. 135–142. IEEE (2017)
2. Hamroun, M., Lajmi, S., Nicolas, H., Amous, I.: Arabic text-based video indexing and retrieval system enhanced by semantic content and relevance feedback. In: 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). pp. 1–8. IEEE (2019)
3. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1580–1589 (2020)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
5. Jain, M., Mathew, M., Jawahar, C.: Unconstrained scene text and video text recognition for arabic script. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). pp. 26–30. IEEE (2017)
6. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1156–1160. IEEE (2015)
7. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: AAAI. pp. 11474–11481 (2020)
8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
9. Lu, T., Palaiahnakote, S., Tan, C.L., Liu, W.: Video text detection. Springer (2014)
10. Mirza, A., Zeshan, O., Atif, M., Siddiqi, I.: Detection and recognition of cursive text from video frames. EURASIP Journal on Image and Video Processing **2020**(1), 1–19 (2020)

11. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., et al.: Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1582–1587. IEEE (2019)
12. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 658–666 (2019)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
14. Tan, M., Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv e-prints arXiv:1905.11946 (2019)
15. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 9627–9636 (2019)
16. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9336–9345 (2019)
17. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
18. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
19. Zayene, O., Hajjej, N., Touj, S.M., Mansour, S.B., Hennebert, J., Ingold, R., Amara, N.E.B.: Icpr2016 contest on arabic text detection and recognition in video frames-activcomp. In: Pattern Recognition (ICPR), 2016 23rd International Conference on. pp. 187–191. IEEE (2016)
20. Zayene, O., Hennebert, J., Ingold, R., Amara, N.E.B.: Icdar2017 competition on arabic text detection and recognition in multi-resolution video frames. In: 2017 International Conference on Document Analysis and Recognition. pp. 1460–1465. IEEE (2017)
21. Zayene, O., Hennebert, J., Touj, S.M., Ingold, R., Amara, N.E.B.: A dataset for arabic text detection, tracking and recognition in news videos-activ. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 996–1000. IEEE (2015)
22. Zayene, O., Touj, S.M., Hennebert, J., Ingold, R., Amara, N.E.B.: Data, protocol and algorithms for performance evaluation of text detection in arabic news video. In: Advanced Technologies for Signal and Image Processing (ATSIP), 2016 2nd International Conference on. pp. 258–263. IEEE (2016)
23. Zayene, O., Touj, S.M., Hennebert, J., Ingold, R., Amara, N.E.B.: Open datasets and tools for arabic text detection and recognition in news video frames. Journal of Imaging **4**(2), 32 (2018)
24. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9759–9768 (2020)