ICDAR2017 Competition on Arabic Text Detection and Recognition in Multi-resolution Video Frames

Oussama Zayene*†, Jean Hennebert‡, Rolf Ingold† and Najoua Essoukri BenAmara*
* LATIS Lab, National Engineering School of Sousse (Eniso), University of Sousse, Sousse, Tunisia
Email: najoua.benamara@eniso.rnu.tn
† DIVA group, Department of Informatics, University of Fribourg (Unifr) Fribourg, Switzerland
Email: {oussama.zayene, rolf.ingold}@unifr.ch
‡Institute of Complex Systems, HES-SO, University of Applied Science Western Switzerland, Switzerland
Email: jean.hennebert@hefr.ch

***Abstract*—This paper describes the multi-resolution Arabic Text detection and recognition in Video Competition—AcTiVComp held in the context of the 14th International Conference on Document Analysis and Recognition (ICDAR'2017), during November 9-15, 2017, in Kyoto, Japan. The main objective of this competition is to evaluate the performance of participants' algorithms for automatically detecting and recognizing Arabic texts in video frames using the freely available Arabic-Text-in-Video (AcTiV) dataset. A first edition was held in the framework of the 23rd International Conference on Pattern Recognition (ICPR'2016). Three groups with five systems are participating to the second edition of AcTiVComp. These systems are tested in a blind manner on a closed-subset of the AcTiV database, which is unknown to all participants. In addition to the experimental setup and observed results, we also provide a short description of the participating groups and their systems.**

***Keywords*— Arabic Text Detection, Arabic Text Recognition, Video-OCR, AcTiV dataset, ICDAR competition**

I. Introduction

Despite the presence of several Arabic news channels with very high viewing rates in the Arabic world and outside of it there are only very few researches addressing the problem of Arabic video text detection and recognition [2], [10], [20]. Actually, we need to extract embedded texts from video content in order to provide powerful semantic cues for automatic broadcast annotation and large archive managing. Texts in videos are more difficult to extract and recognize than texts in scanned documents. This is 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 past ten years, Arabic handwriting and printed text recognition systems have achieved considerable improvements [7], [9]. A lot of progress of such systems has been triggered thanks to the availability of benchmarking databases [1], [8], [11], [16] and the organization of competitions [6], [15], [17]. While printed and handwritten tasks are rather well covered, organized competitions [21] and publicly available datasets for Arabic text detection and recognition in videos [19], [22] are limited to little work. So far, most of the existing systems for overlaid text detection and recognition in Arabic news video [2], [10] are tested on private datasets with non-uniform evaluation protocols, which make objective comparison and scientific benchmarking rather impractical. The goals of this competition are i) to evaluate the performance of participating systems for automatically detecting and recognizing artificial texts in Arabic multi-resolution video frames and ii) to unify the ground-truth formats, the evaluation protocols and the metrics used in the Arabic Video OCR domain. For these purposes, we provide the participants with a new version of the Arabic-Text-in-Videos (AcTiV) dataset [22], which is freely available for researchers.

This is the second competition of its series with the addition of new stream-resolution (480x360) and data (youTube quality). The first competition was organized in conjunction with ICPR'2016 [21] and received a very encouraging response with a total of four submissions from three different institutions. The best F-score results for the detection protocols p4.4 and p7 did not exceed 0.8. In the recognition task, the best results for the p6.4 protocol did not exceed 0.36 in terms of Line Recognition Rate (LRR). Furthermore, there was no participation for protocol p9, and only one system was submitted for protocol p7. To this end, we are organizing a new edition so as to improve these results, especially for the *channel-free* evaluation protocols. Additionally, this second edition has a special focus on evaluating the multi-resolution impact on the detection and recognition performances. The participating systems will be evaluated in both *channel-dependent* and *channel-free* contexts. For this, new evaluation protocols are defined.

AcTiVComp has been organized by LATIS -Laboratory of Advanced Technology and Intelligent Systems- from the National Engineering School of Sousse, Tunisia and DIVA -Document, Image and Voice Analysis- group, from the University of Fribourg, Switzerland in collaboration with ICoSys -Institute of Complex Systems- from the University of Applied Sciences and Arts, Western Switzerland. In this second edition of AcTiVComp, three groups with five systems are participating to the contest. These systems are tested in a blind manner on the closed-test set of AcTiV, which is unknown to all participants.

In the following, we first present the used datasets in section II. Section III is dedicated to the competition protocols. We

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.238

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

```
<?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>
```

Fig. 2. Part of detection XML file of TunisiaNat1 TV channel

describe the participating systems in section IV. The results are discussed in Section V and section VI provides the conclusion.

II. Competition Datasets

AcTiV was presented in the ICDAR 2015 conference as a first publicly accessible annotated dataset designed to assess the performance of different Arabic Video OCR systems [22]. This database is currently used by several research groups around the world. It was partially used as a benchmark in the first edition of AcTiVComp. The two main challenges addressed by this dataset are: i) the variability of text patterns, e.g. text colors, fonts, sizes and position; and ii) the presence of complex backgrounds with various text-like objects. The current version of this dataset consists of 100 video clips recorded from four different Arabic news channels: TunisiaNat1, France24 Arabic, Russia Today Arabic and AljazeeraHD. AcTiV includes two appropriate datasets: AcTiV-D for detection tasks and AcTiV-R for recognition ones.

A. AcTiV-D

AcTiV-D represents a dataset of non-redundant frames used to measure the performance of single-frame-based methods for detecting text regions in HD/SD frames. These frames

TABLE I
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

were hand-selected with a particular attention to achieve a high diversity in depicted text regions. Figure 1 illustrates examples from AcTiV-D dataset for typical problems in video text detection. The dataset contains a total of 2,557 frames distributed over four sets (one set per 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 I. To test the systems abilities for detecting / locating texts under different situations, the proposed dataset includes some frames that do not contain any text and some others that contain the same text regions but with different backgrounds. The detection ground-truth XML file is provided at the line level for each frame. Figure 2 depicts a part of a ground-truth XML file. One bounding box is described by the element *rectangle*, which contains the rectangle attributes: (x, y) co-ordinates, width and height. This XML file was generated by our semi-automatic annotation framework published in [23].

B. AcTiV-R

AcTiV-R represents a dataset of cropped textline images used to evaluate the performance of Arabic text recognition systems. It consists of 10,415 cropped textline images, 44,583 words and 259,192 characters distributed over four sets (one set per TV channel). As shown in figure 3, AcTiV-R texts are in various fonts and sizes and with different degrees of background complexity. The recognition ground-truth is provided at the line level for each text image. Figure 4 depicts an example of a ground-truth XML file and its corresponding textline image. During the annotation process, 165 character shapes were considered, including 10 digits and 12 punctuation marks. More details about the protocols and statistics of the used dataset are given in Table II. The recognition ground-truth files are provided at the line level for each textline image. The XML file is composed of two principal markup sections: ArabicTranscription and LatinTranscription. In order 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. The transcription sample نشرة الأخبَار is as: *Nuun_B Shiin_M Raa_E TaaaClosed_I Space Alif_I Laam_EHamzaAboveAlif_E Xaa_B Baa_M Alif_E Raa_I.*

TABLE II
RECOGNITION DATASET AND EVALUATION PROTOCOLS."LNS" AND "WDS" RESPECTIVELY DENOTE "LINES" AND "WORDS"

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
	Russia Today arabic	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 Youtube	-	-	-	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

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

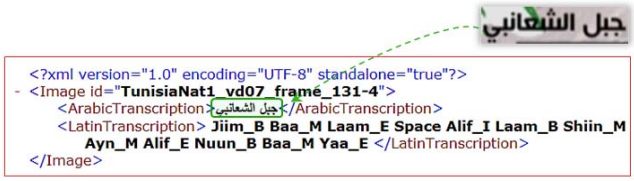

Fig. 4. Recognition ground-truth file and its corresponding textline image

III. PERFORMANCE EVALUATION

Different evaluation metrics and protocols have been proposed for text detection and recognition research fields. In this edition as in the first one, we adopt the same metrics used in several ICDAR competitions [4], [5], [13], [17]. In this section, we describe the used metrics and evaluation protocols for the detection and recognition tasks, respectively.

A. Detection protocols and metrics

Table I depicts the detection protocols.

- **Protocol 1** aims to measure the performance of single-frame-based methods to localize texts in HD frames.
- **Protocols 4** are similar to protocol 1, differing only in channel resolution. All SD (720x576) channels in our database are targeted by these protocols, which are 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 new added resolution (480x360) for the TunisiaNat1 TV channel. The main idea of this protocol is to train a given system with SD (720x576) data, i.e. protocol p4.3, and test it with different data resolution and quality.
- **Protocol 7** is the generic version of the previous protocols where text detection is evaluated regardless of data quality.

Metrics: The performance of a text detector is evaluated based on precision, recall and F-measure metrics that are defined by equations (1), (2) and (3), respectively.

$$Precision = \frac{\sum_{i=1}^{|D|} matchD(D_i)}{|D|} \quad (1)$$

$$Recall = \frac{\sum_{i=1}^{|G|} matchG(G_i)}{|G|} \quad (2)$$

$$Fmeasure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

where D is the list of detected rectangles, G is the list of ground-truth rectangles, and matchD / matchG are the matching functions, respectively. These measures are calculated utilizing our evaluation tool [24], which takes into account all types of matching cases between G bounding boxes and D ones, i.e. one-to-one, one-to-many and many-to-one matching. The proposed performance metrics are similar to those used in ICDAR 2013 [5] and ICDAR 2015 [4].

B. Recognition protocols and metrics

Table II presents the recognition protocols.

- **Protocol 3** aims to evaluate the performance of OCR systems to recognize texts in HD images.
- **Protocols 6** are similar to protocol 3, differing only in channel resolution. All SD (720x576) channels in our database are targeted by these protocols split in four sub-protocols: three *channel-dependent* protocols (p6.1, p6.2 and p6.3) and a *channel-free* one (p6.4).
- **Protocol 6bis** is dedicated to the new stream-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.

Metrics: The performance measure for the recognition task is based on the LRR and the Word Recognition Rate (WRR) at the line and word levels, respectively, and on the computation of insertion (I), deletion (Dl) and substitution (S) errors at the

level of Character Recognition Rate (CRR), which are defined as follows:

$$CRR = \frac{\#characters - I - S - Dl}{\#characters} \quad (4)$$

$$WRR = \frac{\#words\_correctly\_recognized}{\#words} \quad (5)$$

$$LRR = \frac{\#lines\_correctly\_recognized}{\#lines} \quad (6)$$

IV. Participating Systems

Next, we give brief descriptions of the submitted detection and recognition systems provided by the participating researchers.

A. THDL systems

The THDL systems are submitted by Ruijie Yan, Donglai Xiang, Yaqi Wang, Xuecheng Wang and Liangrui Peng, from the Department of Electronic Engineering, Tsinghua University, Beijing, China.

THDL-Det: For the detection task, the authors present a Convolutional Neural Network (CNN) architecture with a multi-level feature pyramid. It consists of a modified Fully-Convolutional Network (FCN) with a residual connection as a proposal generator and a Fast R-CNN detector with rotation RoI pooling for multi-oriented text detection. First, the input image is fed into an FCN, which predicts a salient map. The latter contains the probability of every pixel belonging to a text region. This map is then binarized at multiple thresholds, and Connected Components (CCs) are extracted. The CCs that break into multiple parts at a higher threshold are selected and their bounding boxes represent region proposals. Next, the features of the region proposals after rotation RoI pooling are input into a Fast R-CNN network that filters non-text regions and regresses the bounding boxes to more accurate positions. Finally, non-maximum suppression is performed to obtain text detection results. The network is pre-trained with publicly available datasets including SynthText. The pre-trained network is fine-tuned with the training set of AcTiV-D. The system for the detection task is implemented using PyTorch.

THDL-Rec: For the recognition task, a Recurrent Neural Network (RNN) with Gated Recurrent Units (GRUs) is put forward to recognize Arabic texts in video [18]. The authors construct a four-layer bi-directional RNN model with a Connectionist Temporal Classification (CTC) output layer. Dropout and sparse training mechanisms are adopted to mitigate overfitting. During the sparse training process, small weights are pruned to gain sparsity in network parameters. The original images are converted into grayscale images and normalized with the same height (e.g. 48 pixels). They also generate some image samples that contain digits as a supplement to the training set, since samples with digits are rare in the competition training set. The system for the recognition task is implemented using TensorFlow.

B. CLS-Det system

The CLS-Det system is submitted for the detection task by Wenhao He, Kun-Ze Chen, Fei Yin, Cheng-Lin Liu from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China.

This method is mainly based on the published work [3] for multi-oriented text detection, and the whole system contains two stages. For the first stage, a fully convolutional-based network is trained with a bi-task output, where one task is a segmentation between text and non-text pixels, and the other task is a direct regression [3] to learn the offsets from a given pixel to the corresponding text boundaries. Consequently, each pixel forms a scored bounding box through both tasks. Taking into account that only horizontal text lines should be detected, the regression task only learns to describe rectangles. For the second stage, as there are long text lines that should be detected, the authors perform line grouping for text line candidates already obtained in the first step. The text line candidates are the scored bounding boxes with high text confidence in the segmentation task. Text line candidates that are horizontally close to each other and that have similar height are grouped together into a text line. After getting text lines, the authors will expand text-line boundaries if they cross text CCs. The used training data is divided into two groups: text data and non-text data. Text data are collected only from the training data provided by the competition. Whereas, non-text data are collected from other benchmarks like ICDAR2013 / ICDAR2015 Scene Text datasets [4], [5] and ICDAR2017 Competition of Reading Chinese Text in the Wild.

C. DCR-Rec system

The DCR-Rec system is submitted for the recognition task by Yanfei Lv, Yichao Wu, Fei Yin and Mingchao Xu, from the NLPR, Chinese Academy of Sciences, China.

This system is based on a deep CNN and a Bi-Directional Long-Short Term Memory (Bi-LSTM) architecture, which are widely used in text recognition and sequence translation. The latter architecture is coupled with a CTC component in order to learn feature sequence labeling without any prior segmentation. Benefiting from the convolution on whole text images, the system is easily trained on millions of text lines on a NVIDIA TITAN X(Pascal) GPU. In the encoder part, the authors process the images through a CNN, and then get high-level features. The CNN model is derived from the well-known VGG net [14]. To increase the generalization performance, dropout and batch normalization methods are applied on some layers. Afterwards, the high-level line features are passed into a Bi-LSTM network. The Bi-LSTMs stacked on top of the CNN architecture can process arbitrary-length line images with a strong capability of capturing contextual information. In the decoder stage, the authors apply a lexicon-free beam-search-based algorithm to transcribe features into a label sequence. A simple N-gram language model is applied and trained on a text corpus containing about 100 million characters. The system is purely trained with a synthetic dataset containing more than 300 classes. The synthetic line images

TABLE III
RESULTS OF DETECTION PROTOCOLS

System/Protocol		P1	P4.1	P4.2	P4.3	P4.3bis	P4.4	P7
THDL-Det	Precision	**0.82**	**0.85**	**0.88**	**0.90**	**0.86**	**0.87**	**0.86**
	Recall	**0.79**	**0.81**	**0.86**	**0.89**	**0.81**	**0.85**	**0.83**
	F-measure	**0.80 (1)**	**0.83 (1)**	**0.87 (1)**	**0.90 (1)**	**0.83 (1)**	**0.86 (1)**	**0.85 (1)**
CLS-Det	Precision	0.36	0.35	0.75	0.71	0.71	0.60	0.54
	Recall	0.69	0.64	0.71	0.77	0.47	0.70	0.70
	F-measure	0.48 (3)	0.45 (3)	0.73 (2)	0.74 (2)	0.57 (3)	0.65 (2)	0.61 (3)
Ds-ATD	Precision	0.68	0.63	0.63	0.54	0.58	0.58	0.59
	Recall	0.72	0.61	0.55	0.69	0.67	0.65	0.67
	F-measure	0.70 (2)	0.62 (2)	0.59 (3)	0.60 (3)	0.62 (2)	0.61 (3)	0.62 (2)

TABLE IV
RESULTS OF RECOGNITION PROTOCOLS

System/Protocol		P3	P6.1	P6.2	P6.3	P6.3bis	P6.4	P9
DCR-Rec	CRR	**99.67**	**97.70**	**98.77**	98.77	**97.30**	98.53	98.81
	WRR	**94.14**	**89.04**	**88.93**	82.11	**81.19**	83.4	86.31
	LRR	**89.69 (1)**	**69.63 (1)**	**74.61 (1)**	58.82 (2)	**65.73 (1)**	68.26 (2)	74.30 (2)
THDL-Rec	CRR	99.25	97.63	98.37	**98.96**	95.87	**98.38**	**98.59**
	WRR	90.72	88.23	84.66	**87**	79.13	**85.7**	**87.05**
	LRR	85.50 (2)	67.54 (2)	69.14 (2)	**78.73 (1)**	52.58 (2)	**71.86 (1)**	**75.70 (1)**

are resized to 256x32 and processed with gray normalization before being fed into the system. The SGD technique is used with a momentum for training. The learning rate is initialized as 1e-4 and reduced one time prior to termination. Training this network takes about 30 hours to converge.

D. DS-ATD system

The DS-ATD system [12] is submitted for the detection task by Zied Selmi, Mohamed Ben Halima and Adel M. Alimi., which are members of the REGIM -REsearch Group on Intelligent Machines- at the National Engineering School of Sfax, Tunisia. The submitted system is composed of two phases, namely the pre-processing step and the application of a CNN model.

First, the input frame is converted into HSV (Hue, Saturation and Value) and its contrast is maximized in order to extract small elements and details. Then a set of morphological operations is applied, i.e the top hat transform. After that, Gaussian blur filter is used to remove noise from this frame using a 5x5 kernel. An adaptive thresholding is applied on the obtained image to eliminate irrelevant regions. The authors use then a hierarchical technique to create a full family hierarchy list that aims to find a curve joining all continuous points having the same color and intensity. Finally, a set of geometric filtering rules are applied to extract the possible various bounding boxes that can be considered as text. To decide whether a given bounding box contains a text or not, the authors integrate, in the second phase, a deep learning architecture represented by a CNN model. A bounding box is retained if and only if the prediction is positive and greater than 0.8, which is the minimum threshold of the score obtained by the classifier.

V. RESULTS AND DISCUSSIONS

We present here the results and rankings of the participating systems. We first discuss the results of the text detection task followed by those of text recognition.

Table III summarizes the detection rates of the submitted systems for the seven evaluation protocols. The number within the parenthesis represents the system rank.

The THDL-Det system outperforms the two other systems in all detection protocols realizing F-score rates ranging from 0.8 to 0.9 for protocol p1 (AljazeeraHD) and protocol p4.3 (TunisiaNat1), respectively. This system provides an effective score of 0.85 for the *channel-free* protocol p7, which implies the generalization ability of such a system and its robustness in detecting text regions regardless data resolution. The CLS-Det system achieves good results for protocols p4.2, p4.3 and p4.4 with an F-score of 0.73, 0.74 and 0.65, respectively. Nevertheless, the scores of the remaining protocols are quite low especially in terms of precision metric. The DS-ATD system results are around 0.61 for all protocols and 0.7 for AljazeeraHD's protocol.

We notice that the participating systems are affected by the image quality in protocol p4.3bis (SD 480x360), with a significant decrease in the F-score metric, except for the DS-ATD system which does not decline but on the contrary, it obtains an F-score that is roughly 2% higher than the one in protocol p4.3 (SD 720x576).

Another interesting observation that can be drawn from the realized results is that all participating systems use a quite similar CNN-based architecture, but differ in how they deal with the original image in the first stage, i.e. proposal-based technique (TH-DL system), pixel-based classification (CLS-Det system) or a set of heuristic pre-processing steps (DS-ATD system). The three latter could have an impact on the use of CNNs in the next stage.

Table IV presents the recognition performances of the submitted systems for the seven evaluation protocols. The best result is marked in bold. The DCR-Rec system shows a

superiority in the p3, p6.1, p6.2 and p6.3bis *channel-depending* protocols realizing a best LRR of 0.89 for HD resolution. The THDL-Rec system performs quite better in the p6.4 and p9 *channel-free* protocols as well as in the p6.3 protocol realizing a best LRR of 0.78 for SD resolution.
It is interesting to note that the obtained results in the global protocol p9, which are around 0.75 in terms of LRR, represent a significant improvement in the Arabic Video OCR field. An other important observation is that both systems use Bi-RNNs but in a different way. The first system applies dropout and sparse training techniques and the second one uses a hybrid RNN-CNN representation.

It can be seen from Tables III and IV that the detection and recognition rates of the winning systems outperform the state-of-the-art performances on this problem. In the previous competition, the highest detection rate of 86% was reported on the TunisiaNat subset of the same database, compared to this competition where the rate is 90%. For the recognition protocols there is an improvement of 8% compared to the highest obtained LRR (p3) of AcTiVComp16.

VI. Conclusions

The second edition of AcTiVComp has attracted three groups for participating and has received five systems for two tasks: text detection and textline recognition. The best results have been yielded by the system of Ruijie Yan et al. (THDL-D) for all detection protocols. For the recognition task, the DCR-Rec system has score best for the *channel-depending* protocols and the THDL-Rec has score quite better for the *channel-free* protocols. 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] Randa IM Elanwar, Mohsen A Rashwan, and Samia A Mashali. Ohasd: the first on-line arabic sentence database handwritten on tablet pc. *Analysis*, 2109:881, 2010.

[2] Mohamed Ben Halima, Adel Alimi, Ana Fernández Vila, et al. Nf-savo: Neuro-fuzzy system for arabic video ocr. *arXiv preprint arXiv:1211.2150*, 2012.

[3] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. *arXiv preprint arXiv:1703.08289*, 2017.

[4] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1156–1160. IEEE, 2015.

[5] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere de las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013.

[6] Monji Kherallah, Najiba Tagougui, Adel M Alimi, Haikal El Abed, and Volker Margner. Online arabic handwriting recognition competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1454–1458. IEEE, 2011.

[7] Liana M Lorigo and Venugopal Govindaraju. Offline arabic handwriting recognition: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):712–724, 2006.

[8] Sabri A Mahmoud, Irfan Ahmad, Wasfi G Al-Khatib, Mohammad Alshayeb, Mohammad Tanvir Parvez, Volker Märgner, and Gernot A Fink. Khatt: An open arabic offline handwritten text database. *Pattern Recognition*, 47(3):1096–1112, 2014.

[9] Volker Märgner and Haikal El Abed. *Guide to OCR for arabic scripts*. Springer, 2012.

[10] Mohieddin Moradi and Saeed Mozaffari. Hybrid approach for farsi/arabic text detection and localisation in video frames. *IET Image Processing*, 7(2):154–164, 2013.

[11] Mario Pechwitz, S Snoussi Maddouri, Volker Märgner, Noureddine Ellouze, Hamid Amiri, et al. Ifn/enit-database of handwritten arabic words. In *Proc. of CIFED*, volume 2, pages 127–136. Citeseer, 2002.

[12] Zied Selmi, Mohamed Ben Halima, and Adel M Alimi. Deep learning system for automatic license plate detection and recognition. In *Document Analysis and Recognition (ICDAR), 2017 14th International Conference on*. IEEE, 2017.

[13] Asif Shahab, Faisal Shafait, and Andreas Dengel. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *2011 international conference on document analysis and recognition*, pages 1491–1496. IEEE, 2011.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Fouad Slimane, Sameh Awaida, Anis Mezghani, Mohammad Tanvir Parvez, Slim Kanoun, Sabri A Mahmoud, and Volker Märgner. Icfhr2014 competition on arabic writer identification using ahtid/mw and khatt databases. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 797–802. IEEE, 2014.

[16] Fouad Slimane, Rolf Ingold, Slim Kanoun, Adel M Alimi, and Jean Hennebert. A new arabic printed text image database and evaluation protocols. In *2009 10th International Conference on Document Analysis and Recognition*, pages 946–950. IEEE, 2009.

[17] Fouad Slimane, Slim Kanoun, Haikal El Abed, Adel M Alimi, Rolf Ingold, and Jean Hennebert. Icdar2013 competition on multi-font and multi-size digitally represented arabic text. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013.

[18] Ruijie Yan, Liangrui Peng, and Guangxiang Bin. Residual recurrent neural network with sparse training for offline arabic handwriting recognition. In *Document Analysis and Recognition (ICDAR), 2017 14th International Conference on*. IEEE, 2017.

[19] Sonia Yousfi, Sid-Ahmed Berrani, and Christophe Garcia. Alif: A dataset for arabic embedded text recognition in tv broadcast. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1221–1225. IEEE, 2015.

[20] Sonia Yousfi, Sid-Ahmed Berrani, and Christophe Garcia. Deep learning and recurrent connectionist-based approaches for arabic text recognition in videos. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1026–1030. IEEE, 2015.

[21] Oussama Zayene, Nadia Hajjej, Sameh Masmoudi Touj, Soumaya Ben Mansour, Jean Hennebert, Rolf Ingold, and Najoua Essoukri Ben Amara. Icpr2016 contest on arabic text detection and recognition in video frames-activcomp. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 187–191. IEEE, 2016.

[22] Oussama Zayene, Jean Hennebert, Sameh Masmoudi Touj, Rolf Ingold, and Najoua Essoukri Ben Amara. A dataset for arabic text detection, tracking and recognition in news videos-activ. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 996–1000. IEEE, 2015.

[23] Oussama Zayene, Sameh Masmoudi Touj, Jean Hennebert, Rolf Ingold, and Najoua Essoukri Ben Amara. Semi-automatic news video annotation framework for arabic text. In *2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2014.

[24] Oussama Zayene, Sameh Masmoudi Touj, Jean Hennebert, Rolf Ingold, and Najoua Essoukri Ben Amara. 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*, pages 258–263. IEEE, 2016.