2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

ICPR2016 Contest on Arabic Text Detection and Recognition in Video Frames−AcTiVComp

Oussama Zayene[1, 2], Nadia Hajjej[1], Sameh Masmoudi Touj[1], Soumaya Ben Mansour[1], Jean Hennebert[2, 3], Rolf Ingold[2] and Najoua Essoukri Ben Amara[1]

[1] *SAGE: Systèmes Avancés en Génie Electrique Research Unit, National Engineering School of Sousse (ENISo), University of Sousse, BP 264 Sousse Erriadh 4023 Tunisia*

[2] *DIVA: Document, Image and Voice Analysis research Group, Department of Informatics University of Fribourg (Unifr), Bd de Pérolles 90, CH-1700 Fribourg, Switzerland*

[3] *ICoSys: Institute of Complex Systems, HES-SO // Fribourg, University of Applied Sciences and Arts, Western Switzerland*

oussema.zayene@unifr.ch, hajjej.nadia@gmail.com, samehmasmouditouj@yahoo.fr, jean.hennebert@hefr.ch, rolf.ingold@unifr.ch, najoua.benamara@eniso.rnu.tn

***Abstract*—This paper describes the AcTiVComp: detection and recognition of Arabic Text in Video competition in conjunction with the 23rd International Conference on Pattern Recognition (ICPR). The main objective of this competition is to evaluate the performance of participants' algorithms to automatically locate and/or recognize overlay text lines in Arabic video frames using the freely available AcTiV dataset. In this first edition of AcTiVComp, four groups with five systems are participating to the competition. In the detection challenge, the systems are compared based on the standard assessment metrics (i.e. recall, precision and F-score). The recognition results evaluation is based on the recognition rates at the character, word and line levels. The systems were tested in a blind manner on the *closed-test set* of the AcTiV dataset which is unknown to all participants. In addition to the test results, we also provide a short description of the participating groups and their systems.**

Keywords- ICPR contest; overlay Arabic text; text detection; text recogntion; VOCR; AcTiV dataset

I. Introduction

Recognition of Arabic texts and analysis/indexing of Arabic documents have become recently a compelling research domain. Several techniques have been proposed in the conventional field of Arabic Optical Character Recognition (AOCR) [1] (e.g., scanned documents). However, few attempts have been made on the development of detection/recognition systems for overlay text in Arabic news videos. Compared to classical documents, text detection and recognition in video frames is more challenging due to the complexity of background (e.g., presence of text-like objects), unknown text size/color, degraded text quality caused by compression artifacts, etc.

Recently, more advanced approaches have been proposed for detection and recognition of text in videos and natural scene images [2, 3] [6-8]. Most of these methods are tested and compared in the context of international competitions (e.g., ICDAR'13 [4], ICDAR'15 [5], etc.), taking advantage of freely available standard datasets [4-8]. Unfortunately, little is known about the behavior and performance of already published detection /recognition systems dedicated to Arabic video texts [9, 13] due to the absence of reliable benchmarking and direct comparison; because very limited effort was spent to develop benchmark datasets. To the best of our knowledge, Arabic Text in Video (AcTiV) dataset [10] is the first publicly accessible annotated dataset designed to assess the performance of Arabic video text detection, tracking and recognition systems. The challenges that are addressed by AcTiV are in text patterns variability (i.e., different colors, unknown sizes/fonts, etc.) and presence of complex background. Another challenging aspect of this dataset is linked to the peculiarities of Arabic text which include non-uniform intra/inter word distances, diacritics, cursive nature of the script, etc.

The main objectives of the Arabic Text in Video Competition− AcTiVComp are i) to overcome these problems by suggesting promising approaches for automatic video text detection and / or recognition and ii) to unify the evaluation protocols / methodology used in the Arabic Video OCR research. AcTiVComp was organized by the SAGE (Systèmes Avancés en Génie Electrique) research unit, from the National Engineering School of Sousse, Tunisia and the DIVA (Document, Image and Voice Analysis) research group, from the University of Fribourg, Switzerland in collaboration with the ICoSys (Institute of Complex Systems) research institute, from the University of Applied Sciences and Arts, Western Switzerland. In this first edition of AcTiVComp, four groups with five systems are participating to the contest. These systems were tested in a blind manner on the *closed-test set* of the AcTiV dataset which is unknown to all participants.

In the following, we first describe the used datasets in Section 2. Section 3 describes the competition protocols. We present participating systems in section 4. Section 5 presents the evaluation results and Section 6 provides conclusions.

II. Contest Datasets

AcTiV [1] is the first publicly accessible annotated dataset designed to assess the performance of different Arabic Video-OCR systems. The two main challenges addressed by AcTiV dataset are: i) Text patterns variability e.g., text colors, fonts, sizes, position, etc. ii) Presence of complex background with

[1] http://tc11.cvc.uab.es/datasets/AcTiV_1

various text-like objects and anti-aliasing/compression artifacts. The dataset includes more than 150 video sequences collected from 4 different Arabic news channels: TunisiaNat1, France24 Arabic, Russia Today Arabic and AljazeeraHD (see figure 1). Two types of video stream were chosen: Standard-Definition (SD) and High-Definition (HD). The proposed corpus includes two datasets, dedicated to the detection and recognition tasks, which will be detailed in the following sections.

A. AcTiV-D dataset

AcTiV-D (D for Detection) represents a subset of non-redundant frames collected from AcTiV dataset and used to measure the performance of single-frame based methods to locate text regions in still HD/SD frames. AcTiV-D consists of 1843 frames distributed over four sets (one set per channel). Every set includes two sub-sets: trainingFiles and testFiles.

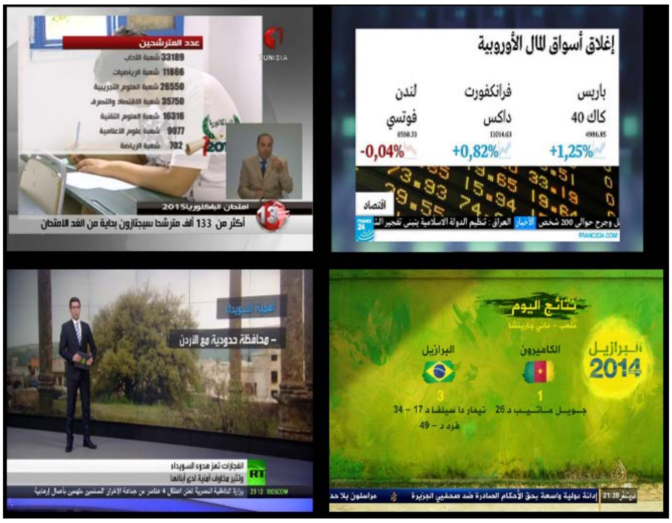
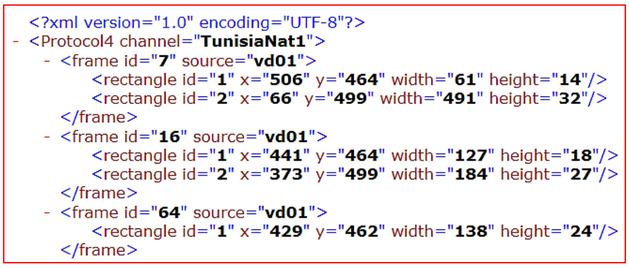


Figure 1. Typical video frames from AcTiV-D dataset. Top Sub-figures: examples of TunisiaNat1 and France24 frames. Bottom Sub-figures: examples of RussiaToday and Aljazeera HD frames.

The Detection groundtruth xml file is provided at the line level for each frame. Figure 2 depicts a part of a groundtruth xml file. One bounding box is described by the element <Rectangle> which contains the rectangle's attributes: (x, y) coordinates, width and height. This xml file was generated by our Semi-automatic annotation framework published in [11].



```
  <?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>
```

Figure 2. A part of the detection xml file of TunisiaNat1TV channel

B. AcTiV-R dataset

A sub-dataset of cropped text images, named AcTiV-R, was created from AcTiV dataset and used to evaluate the performance of Arabic text recognition systems. As shown in figure 3 AcTiV-R texts are in five different fonts with various sizes and colors. AcTiV-R consists of 8757 textline images (40,749 words) distributed over four sets (one set per channel). Every set includes two sub-sets: *trainingFiles* and *testFiles*.

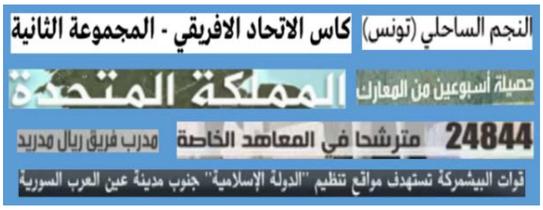
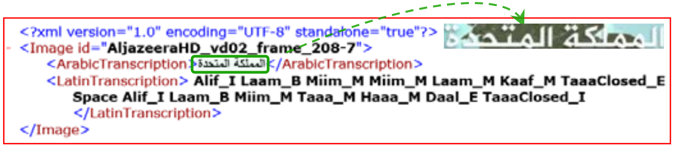

النجم الساحلي (تونس)
كاس الاتحاد الافريقي - المجموعة الثانية
حصيلة أسبوعين من المعارك
المملكة المتحدة
24844 مترشحا في المعاهد الخاصة
مدرب فريق ريال مدريد
قوات البيشمركة تستهدف مواقع تنظيم "الدولة الإسلامية" جنوب مدينة عين العرب السورية

Figure 3. Example of some text images from the AcTiV-R dataset

The Recognition groundtruth files are provided at the line level for each textline image. The xml file is composed of two principal markups sections: ArabicTranscription and LatinTranscription. In order to have an easily accessible representation of Arabic text, it is transformed into a set of labels with a suffix that refers to the letter's position in the word (_B: Begin, _M: Middle, _E: End and _I: Isolate). Figure 4 depicts an example of a groundtruth xml file.



```
  <?xml version="1.0" encoding="UTF-8" standalone="true"?>
- <Image id="AljazeeraHD_vd02_frame_208-7">
      <ArabicTranscription>المملكة المتحدة</ArabicTranscription>
      <LatinTranscription> Alif_I Laam_B Miim_M Miim_M Laam_M Kaaf_M TaaaClosed_E
         Space Alif_I Laam_B Miim_M Taaa_M Haaa_M Daal_E TaaaClosed_I
         </LatinTranscription>
  </Image>
```

Figure 4. A recognition groundtruth file and its corresponding textline image

To evaluate the performance of participating systems in a blind manner we dedicated a closed test datasets for the detection task as well as for the recognition one, called closed-test set. The statistics of these datasets are shown in Table I. The datasets include 413 frames, 956 textline images, 4488 words and 25785 characters.

TABLE I. STATISTICS OF THE CLOSED TEST DATASETS

Resolution	TV channel	AcTiV-D *closed-test set*	AcTiV-R *closed-test set*		
		#frames	**#textline**	**#word**	**#character**
HD (1920x1080)	AljazeeraHD	103	262	1082	6283
SD (720x576)	France 24	104	217	854	4600
	Russia Today	100	256	1598	9305
	TunisiaNat1	106	221	954	5597

III. PERFORMANCE EVALUATION

A set of eleven evaluation protocols was proposed in [10] taking advantage of the variability in data content in terms of text sizes, fonts, colors and background complexity. The participating methods are evaluated under the same experimental conditions and protocols. In this edition, we focus only on the protocols dedicated to the detection and recognition tasks, as depicted in tables II and III respectively.

A. Detection protocols and metrics

Protocol 1 aims to measure the performance of single-frame based methods to localize text regions in still HD images.

Protocol 4 is similar to protocols 1, differing only by the channel resolution. All SD channels in our database can be targeted by this protocol.

Protocol 7 is the generic version of the previous protocols where text detection is evaluated regardless of data quality.

TABLE II. DETECTION EVALUATION PROTOCOLS

Protocol	Resolution	TV Channel		Task
1	HD	AljazeeraHD		D
4	SD	4.1	France 24	D
		4.2	Russia Today	
		4.3	TunisiaNat1	
		4.4	All SD channels	
7	-	All		D

Metrics: The performance of the submitted text detectors has been evaluated based on precision, recall and F-measure metrics using our evaluation tool [12]. This tool takes into account all types of matching cases between groundtruth bounding boxes and detected 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'15 [5].

B. Recognition protocols and metrics

Protocol 3 aims to evaluate the performance of OCR systems to recognize texts in HD frames.

Protocol 6 is similar to protocols 3, differing only by the channel resolution. All SD channels in our database can be targeted by this protocol.

Protocol 9 is the generic version of the previous ones where text recognition is evaluated independently to data quality.

TABLE III. RECOGNITION EVALUATION PROTOCOLS

Protocol	Resolution	TV Channel		Task
3	HD	AljazeeraHD		R
6	SD	6.1	France 24	R
		6.2	Russia Today	
		6.3	TunisiaNat1	
		6.4	All SD channels	
9	-	All		R

Metrics: The performance measure for the recognition task is based on the Recognition Rate at the Line level (LRR) and on the insertion, deletion and substitution errors at the word and character levels.

IV. SUBMITTED SYSTEMS

We invited groups participating to this contest to adapt their systems to the AcTiV dataset and to send us executable programs of their systems. This section gives a short description of the participants' systems.

A. ATD-CH System

The ATD-CH (Arabic Text Detection based on Color Homogeneity) System was submitted for task 1 (detection challenge) by Houda Gaddour, Slim Kanoun and Nicole Vincent in the context of a joint collaboration between the MIRACL "Multimedia, InfoRmation systems and Advanced Computing Laboratory", University of Sfax, Tunisia and the LIPADE laboratory "Laboratoire d'Informatique Paris Descartes", University of Paris Descartes, France. The basic idea of this approach [13] is the consistency of the text color that distinguishes it from the foreground and from the other objects in the same image. Based on the MSER (Maximally Stable Extremal Regions) idea and instead of relying on a range of unique thresholds, a range of pairs of thresholds for each channel in the RGB color space is calculated using k-means algorithm, in order to generate a set of binary maps. This range constructs a set of binary images each belongs to a color interval from [S1, S2]. For all connected components (*CC*) of each binary map, a first filtering is applied according to a stability criterion of the embedded texts to extract candidate components. This criterion is applied for each connected component CC_{0i} in order to test its evolution by varying the two extremities of the color interval [S1, S2] either increasing or reducing it. For a text candidate, the CC surface remains relatively stable since the text is well contrasted compared to the rest of the image. For a non-text component, in most cases there will be a great variation in the surface of the component. The processing function (1) is reflected as follows:

$$\begin{aligned}&Surface\ (CC_{0i}) - Surface(CC_{1i}) < \varepsilon\ \ \&\&\\ &Surface\ (CC_{0i}) + Surface(CC_{2i}) < \varepsilon \end{aligned} \qquad (1)$$

Where *ε:* is the area stability threshold.

CC_{0i} is extracted from the binary image defined by two thresholds [*S1, S2*].

CC_{1i} is extracted from the binary image defined by two thresholds [*S1+ y, S2 - y*] where *y* it's a level for reducing.

CC_{2i} is extracted from the binary image defined by two thresholds [*S1- y, S2 + y*] where *y* it's a level for increasing.

Then, a second filtering process is applied to filter out CCs that are unlikely parts of texts considering the specificities of Arabic script. This process is based on a set of statistical and geometric rules such as Fourier Descriptors, text baseline and ligatures, aspect ratio, etc. Finally a merging process of the remaining components is performed to form textlines.

B. FM-AVTD System

The FM-AVTD (Fast MSER-based Method for Arabic Video Text Detection) System was submitted for task 1 by Yang Xue-hang from the NLPR "National Laboratory of Pattern Recognition", Institute of Automation, Chinese Academy of Sciences, China. The submitted system is based on MSER algorithm with some modifications considering the difference between Arabic and English. First, each input frame is converted into gray map, and Extremal Regions (ERs) are extracted on gray channel and on its reverse channel. Usually more than 1k regions are obtained for each frame, with many overlapping ERs. In order to overcome the repeating components problem and to reduce the number of ERs, a simple suppression method similar to the one in [14] is performed using the following measure (2) that estimates the overlap between ERs based on the hierarchy of the ER tree:

$$Ov(R_t\,, R_{t-k}) = \frac{|R_t|}{|R_{t-k}|} \qquad (2)$$

Where R_{t-k} is the parent of R_t in the ER tree. For each node R_t, the number of overlaps (i.e. with R_{t-k} for all k) n_o is computed such that $ov(R_t, R_{t-k}) > 0.7$. Among the overlapping ERs, those such that $n_o < 3$ are discarded and those with the largest bounding box area are selected. Among the obtained ERs some of them are parts of one character and should be merged into nearby ERs to form one single bounding box. After suppression and preliminary merging steps, the number of ERs is reduced to ten percent, but there are still many noise ERs. Considering the cursive nature of Arabic script, the aspect ratio of Arabic characters varies considerably. Thus, the survived ERs are divided in two sets: *largeAR-set* if their aspect ratio is higher than T_2 (empirically fixed to 1.5), otherwise *smallAR-set*. For each set an AdaBoost classifier is trained, using gray level features, to classify the previously obtained ERs as text and non-text, after resizing them to 16x32(for *largeAR-set*) or 16x16(for *smallAR-set*). After the classification step, nearby ERs are grouped into text lines according to their color similarity, horizontal distance, vertical overlap, height similarity, etc. But it's often the case that the trained classifier misses some positive candidates. Therefore the non-positive ER list is iteratively traversed in order to relabel the ERs as positive if they have similar attributes as described above.

C. D/R-ATVF Systems

The D/R-ATVF (Detection and Recognition of Arabic Text in Video Frames) Systems were submitted by Seiya Iwata, Wataru Ohyama, Tetsushi Wakabayashi and Fumitaka Kimura, from the Graduate School of Engineering, Mie University, Japan. The authors have developed an end-to-end system for Arabic text recognition in video frames [15]. The system was modified to output the enclosing rectangles of detected and recognized text lines. The modified system was submitted as *ProposedDet* for task 1. The end-to-end system itself was submitted as *ProposedRec* for task 2.

Textline detection system: the input image is first converted to a gray scale image and binarized by thresholding operation. The threshold is determined by Otsu's method. Then, CCs are extracted from the binary image by region labeling algorithm, and CCs having width or height greater than predefined thresholds are eliminated as non-text line components. Text lines are detected by means of vertical profile analysis. In order to improve the separability of lines 1-dimensional difference of Gaussian filter (DOG) is applied to the obtained vertical profile map. Black pixels in positive region of the DOG filtered vertical profile are extracted as text lines. Remaining black pixels in the negative region are added to its nearest text lines by fixed neighborhood classification rule. Finally to remove false lines the average of the eccentricity e of a CC is calculated for all components in the textline (3). The line is removed if the average of e is less than 30.

$$e = \frac{perimeter^2}{area} \qquad (3)$$

Word recognition system: After segmenting textlines into words using a space detection algorithm (which consists of classifying the horizontal gaps between CCs as *between_word* gap and *within_word* gap) the word recognition step is performed as follows: the characters are first over segmented and then merged into each character in the process of character recognition [16]. The segmentation points of characters are detected through local extrema analysis at the upper contour of the word image. The word image is physically split at each segmentation points into a set of primitive segments. In order to merge these primitive segments into characters the dynamic programming (DP) is applied to maximize the total likelihood of characters. The likelihood is calculated by the modified quadratic discriminant function (MQDF) using 64-dimensional feature vector of chain code histogram [17]. If the average of the character likelihoods in a word is less than a threshold the word is discarded as a false word. The threshold works as a parameter to trade off the recall and precision of the words.

D. ATR-SID Systems

The ATR-SID (Arabic Textlines Recognition) System was submitted by Soumaya Essefi, Oussama Zayene and Sameh Masmoudi Touj members of the SID "Signal, Image and Document" Team from the SAGE research unit, at the National Engineering School of Sousse, Tunisia. The proposed system [2] [19] presents an implicit segmentation-based recognition technique (i.e. no prior required segmentation of words into characters) using a Recurrent Neural Network (RNN) architecture. This technique relies specifically on a Multi-Dimensional Long Short Term Memory (MDLSTM) with Connectionist Temporal Classification (CTC) output layer. The proposed network is composed of three levels: an input layer, three hidden recurrent layers and an output layer. The hidden layers are MDLSTM that respectively have 2, 10, and 50 cells and separated by feedforward layers with 6 and 20 cells. In fact, the authors have created a hierarchical structure HSRNNs [20] by repeatedly composing MDLSTM layers with feedforward layers. Firstly, the image is divided into input pixel blocks with a size of 2x4 (for task 6.2 and 6.3) and 1x4 (for task 6.1 and 3), each of which is presented to the first MDLSTM layer as a feature vector of pixel intensities. These vectors are then scanned by four MDLSTM layers in different directions. Finally, the activations of the MDLSTM layers are collected into blocks having a size of 2x4 or 1x4. These blocks are given as input to a feedforward layer which can be seen as a subsampling step with trainable weights, in which the activations are summed and squashed by the hyperbolic tangent (*tanh*) function. The final level is the CTC output layer which labels the sequences of textlines. This layer has n cells, where n is the number of classes (n-1 alphabet characters and one node for the blank output). It trains the network to estimate the conditional probabilities of the possible labelling given the input sequences. The output activations are normalised at each timestep with the *softmax* activation function. A separate network has been trained for each TV channels of the reference protocols. All input images have been scaled to common heights and converted to gray-scale (SD images) for training.

V. Results And Discussions

The submitted systems were evaluated on the closed-test set. Each system loads the test samples from hard disk and output the detection/recognition results in a specified format file [18]. The systems can be categorized in two groups

[2] That we declared here "out of competition" for sake of integrity as some members are part of the organizing committee.

depending on the operating system: 3 systems were developed under Linux: D-ATVF, R-ATVF and ATR-SID and 2 systems under Windows environment: ATD-CH and FM-AVTD.

Table IV presents all system results of the detection protocols (task 1) in term of precision, recall and f-measure. The best result is marked in bold. The FM-AVTD system scores best in all protocols (except 4.2): the 4.1 and 4.3 channel-depending protocols and the 4.4 channel-free protocol (resembling all SD channels). The D-ATVF system performs well for all SD protocols except the 4.1 one. However its current version is incompatible with HD resolution. The ATD-CH system has strong fragmentation and miss detection tendency as shown by their obtained numerical results specifically for the precision values in Table IV.

TABLE IV. RESULTS OF DETECTION PROTOCOLS

Protocol System		1	4.1	4.2	4.3	4.4	7
ATD-CH	**Precision** **Recall** **F-measure**	0.40 0.50 0.45	0.47 0.61 0.54	0.35 0.42 0.38	0.34 0.49 0.41		
FM-AVTD	**Precision** **Recall** **F-measure**	**0.71** **0.73** **0.72**	**0.79** **0.79** **0.79**	0.81 0.78 0.79	0.84 **0.88** **0.86**	**0.81** **0.81** **0.81**	0.79 0.79 0.79
D-ATVF	**Precision** **Recall** **F-measure**		0.44 0.44 0.44	**0.83** **0.80** **0.81**	**0.85** 0.85 0.85	0.7 0.7 0.7	

Table V presents the results of task 2 in term of character, word and line recognition rates. The best result is marked in bold. The ATR-SID RNN-based systems have shown superiority in the 6.1, 6.2 and 6.3 channel-depending protocols specifically for the LRR criterion. The R-ATVF system performs better in the 6.4 channel-free protocol.

TABLE V. RESULTS OF RECOGNITION PROTOCOLS

Protocol System		3	6.1	6.2	6.3	6.4	9
ATR-SID	**CRR** **WRR** **LRR**	0.98 0.92 0.81	**0.95** **0.82** **0.57**	**0.96** **0.78** **0.43**	**0.96** **0.81** **0.65**	0.7 0.59 0.3	
R-ATVF	**CRR** **WRR** **LRR**		0.94 0.72 0.45	**0.96** 0.75 0.35	**0.96** **0.81** 0.53	**0.95** **0.73** **0.36**	

VI. CONCLUSION

Four groups presenting five systems have participated at this first edition of the AcTiVComp contest using the AcTiV dataset. Three for text detection task and two for textline recognition task. The best results were yielded by the FM-AVTD systems for the protocols 1, 4.1, 4.3 and 4.4 and by the D-ATVF system for protocol 4.2 in the detection task. In the recognition task, the best results were yielded by the ATR-SID systems for the protocols 6.1, 6.2 and 6.3 and by the R-AVTF system for protocol 6.4.

The contest results show that there is still room for improvement in both detection and recognition of Arabic video text. We look forward to have more participants in the future editions of AcTiVComp and more researchers joining the challenging research topic of Arabic video text detection and recognition.

REFERENCES

[1] V. Märgner and H. El Abed, "Guide to OCR for Arabic Scripts" (book), Springer, 2012.

[2] L. Tong, P.Shivakumara, T. Chew Lim and L. Wenyin., "Video Text Detection" (book), Advances in Computer Vision and Pattern Recognition (ACVPR), 2014.

[3] Q. Ye and D. Doermann. "Text detection and recognition in imagery: A survey". IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI), November 2014.

[4] D. Karatzas et al., "ICDAR 2013 Robust Reading Competition", in Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), August 2013.

[5] D. Karatzas et al., "ICDAR 2015 Robust Reading Competition", in Proc. of ICDAR, August 2015.

[6] K. Wang and S. Belongie, "Word Spotting in the Wild", in Proc. of the 11th European Conference on Computer Vision (ECCV), September 2010.

[7] C. Yao, X. Bai, W. Liu, Y. Ma, Z. Tu, "Detecting texts of arbitrary orientations in natural images", in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2012.

[8] S. Lee, M. S. Cho, K. Jung, and J. Hyung Kim, "Scene Text Extraction with Edge Constraint and Text Collinearity", in Proc. of the International Conference of Pattern Recognition (ICPR), August 2010.

[9] S. Yousfi, S.Berrani, C. Garcia, "Arabic text detection in videos using neural and boosting-based approaches: Application to video indexing", in Proc. of the IEEE International Conference on Image Processing (ICIP), October 2014.

[10] O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, and N. Essoukri Ben Amara, "A dataset for Arabic text detection, tracking and recognition in news videos- AcTiV", in Proc. of ICDAR, August 2015.

[11] O. Zayene, S. M. Touj, J. Hennebert, R. Ingold and N. Essoukri Ben Amara "Semi-Automatic News Video Annotation Framework for Arabic Text", in Proc. of the 4th International Conference Image Processing Theory, Tools and Applications (IPTA), October 2014.

[12] O. Zayene, S. M. Touj, J. Hennebert, R. Ingold and N. Essoukri Ben Amara, "Data, Protocol and Algorithms for Performance Evaluation of text detection in Arabic news Video", in Proc. of the 2nd International Conference on Advanced Technologies for Signal & Image Processing (ATSIP'16), March 2016.

[13] H. Gaddour, S. Kanoun, N. Vincent, "A New Method for Arabic Text Detection in Natural Scene Image Based on the Color Homogeneity", in in Proc. of the 7th International Conference on Image and Signal Processing (ICISP), May 2016.

[14] H. Cho, M. Sung and B. Jun, "Canny Text Detector: Fast and Robust Scene Text Localization Algorithm", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[15] S. Iwata, W. Ohyama, T. Wakabayashi and F. Kimuram, "Recognition and Transition Frame Detection of Arabic News Captions for Video Retrieval", in Proc. of ICPR, December 2016. (to appear)

[16] F. Kimura, M. Shridhar, and Z. Chen, "Improvements of a Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words", in Proc. of ICDAR, October 1993.

[17] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, "Modified quadratic discriminant functions and the application to Chinese character recognition", IEEE Trans. PAMI, January 1987.

[18] http://diuf.unifr.ch/diva/AcTiVComp/run.html

[19] S. Essefi, "Contribution à la reconnaissance des textes Arabes dans les journaux-télévisés", Master's thesis, Ecole Nationale d'Ingenieurs de Sousse (ENISo), 2016.

[20] A. Graves, "Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks." (book chapter), Springer, 2012, pp. 297–313.