ARASTI: A Database for Arabic Scene Text Recognition

Maroua Tounsi*, Ikram Moalla*" and Adel M. Alimi*
*REsearch Groups in Intelligent Machines (REGIM-Lab), ENIS-Sfax, Tunisia
"Al Baha University, Saudi Arabia

Emails: {tounsi.maroua, ikram.moalla, adel.alimi} @ieee.org

Abstract— **Text in natural scenes provides many information for peoples and presents an essential tool to interact with their environment. Therefore, recognizing text existing in camera-captured images has become an important issue for many researches in the last decades. Currently, there isn't any available dataset of Arabic script text images in the wild. Since our aim is to help the research community in standardizing the evaluation of scene Arabic text recognition, we present in this paper a database of images of Arabic Scene Text, segmented scene Arabic words and segmented scene Arabic characters. We call this dataset ARASTI (ARAbic Scene Text Image). This database contains diverse natural scenes images captured at varying weather, lighting and perspective conditions. Moreover, characters and words are also segmented from the original images and stored individually. We obtain 1687 images, 1280 segmented scene Arabic words and 2093 scene Arabic character images. Compared to public datasets of scene text images in other languages like ICDAR03, Chars74K, *etc.*, ARASTI contains a competitive number of images to these databases already published which proves that it can be used as a benchmark.**

Keywords— Arabic scene text, ARASTI Database, Character recognition.

I. INTRODUCTION

Scene character recognition is an imporant step in the process of reading text in the written form. Indeed, text existing all around us presents many informations for peoples. However, tourists in foreign countries are unable to understand what indicate text on shop names, product advertisements, posters, etc. when they are unfamiliar with the foreign language of the visited country.

Today, and with the increase of productivity and profitability with mobile technology, many mobile applications require scene text recognition like for example mobile translators which read text in the wild and translates it into the native language in real time or a vision-based navigation and driving assistant aplications, *etc.* This trend is approved by the important increase of related work [1], [2], [3], [4], [5], [6], [7], [8] in these recent years. Attractive progresses and improvements have been achieved, mainly driven by the competitions and public datasets in this domain, such as the ICDAR Robust Reading competitions [9], [10], [12], [13], MSRA-TD500 [4], SVT [3], Chars74K [11] and IIIT-5K Word [14].

In recent years, many researches had been interested on text recognition in natural scenes for a diversity of languages. Althought, Scene Arabic text recognition problem is not yet well solved due to non-existence of database of available Arabic script text images in the wild and the complex shape nature of Arabic characters depending on their position within a word (beginning, isolated, middle, and end).

In another hand, more than 300 million people in the world speak Arabic. In the last decades, most of researchers in Arabic text recognition have been focused on recognition of scanned off-line printed and handwriting documents and they have developed several benchmark databases and therefore to be able to compare their systems. However, there was not any focus on developing databased and benchmarks for recognition chracaters in natural scenes.

In this paper, we have initiated the development of a real database of images of text in diverse natural scenes captured at varying weather, lighting and perspective conditions for Arabic scene text recognition, called ARASTI (Arabic Scene Text Image Database).

To our knowledge, ours is the first database for Arabic text recognition in real-world.

The paper is organized as follows: In the next section, we describe some related work. In section 3, we present a brief description of Arabic script characteristics. In section 4, we present a description of our database. In the last section, we describe our various prospects and future work.

978-1-5090-6628-5/17/$31.00 ©2017 IEEE

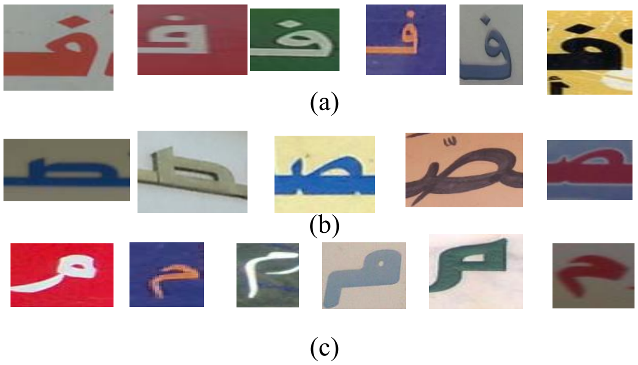


Fig. 1. Some samples for 3 characters presenting different fonts: (a): letter 'ـف', (b) letter 'ـصـ', and (c) letter 'ـم'.

II. RELATED WORK: SCENE TEXT DATASETS

In this section, we describe the several existing scene text database. These are several scene character datasets for different languages including for sets in English, one set in Bengali, one set in Devanagari and one set in Kannada. Some samples of these datasets are presented in fig. 2.

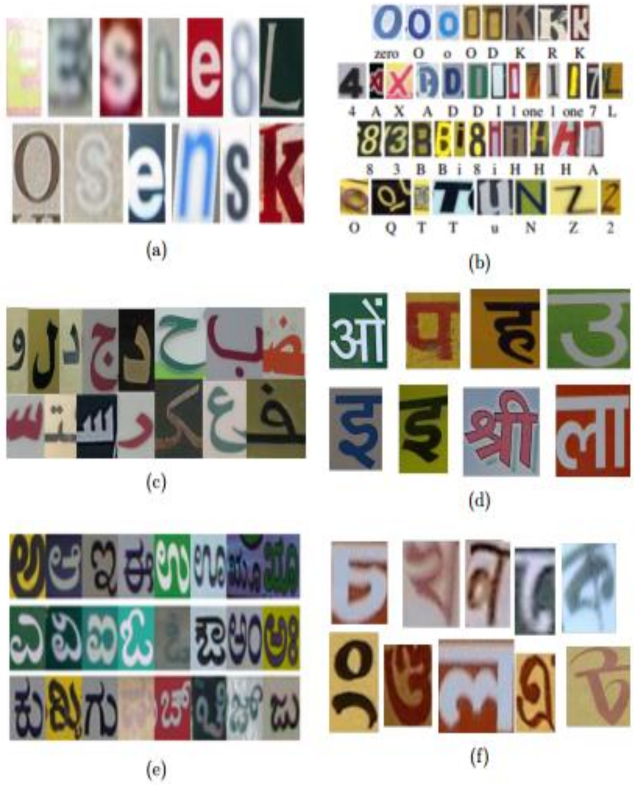


Fig 2. Samples of scene characters of different languages: (a) English scene characters from ICDAR2003 [10], (b) English scene characters from Chars74K [11], (c) Arabic scene characters from images captured in Tunisia, (d) Devangari scene characters from DSIW-3K [16], (e) Kannada scene characters from Chars74K [11], and (f) Bengali scene characters from images captured in West Bengal, India [15] .

B. English scene text dataset

The most widely used in the community is ICDAR 2003 Robust Reading dataset. It contains 509 scene text images.

There is also Chars74K dataset proposed by Campos et al in [11]. It is interested on the recognition of English and Kannada characters in natural scenes images. It contains 1922 images taken from hoardings and advertisements and sign boards.

Another dataset is The Street View Text (SVT) [3] dataset, which is composed by 350 images taken from Google Street View.

B. Bengali scene text dataset

Bengali script is used by Bengali and a few other languages of the eastern part of South Asia. It is the sixth most used script in the world and it is the official language of Bangladesh and India. Bengali alphabet consist on 50 basic characters: 11 vowels and 39 consonants. In [15], the authors proposed a Bengali dataset containing 15250 scene characters (including numerals). These characters are extracted from 260 outdoor scene images captured from the streets of West Bengal.

D. *Devanagari scene text dataset*

This database of the Devanagari characters DSIW-3K [16] is collected from pictures of signboards, hoardings and advertisements in streets, shopping areas, roadside signs, *etc*. Characters are manually extracted from these pictures.

E. Kannada scene text dataset

Kannada is an ancient Dravidian language and it has a history of more than two thousand years. It has 49 basic characters in its alphabets, but consonants and vowels can be combined to give more than 600 distinct classes [11].

The Chars74K Kannada dataset contains a total of 3345 characters were extracted from a set of 1922 photographed images.

We compare these current datasets by statistics on the number of images, and the number of characters. Statistics are included in table I.

Compared to datasets of scene text images, ARASTI contains a competitive number of images to databases already published, which proves that it can be used as a benchmark.

Until now, existing public datasets for Arabic text recognition are limited only to handwritten, printed scanned documents or synthetic texts.

TABLE I. COMPARAISON OF NATURAL IMAGE TEXT RECOGNITION DATASETS.

Dataset	Original Text	Cropped Words	Segmented Characters
ICDAR 2003	509	999	11615
Chars74K-English &Kannada	312	-	3345
Street View Text	350	904	-
IIIT-5K	-	5000	-
Bengali dataset	260	-	15250
ARASTI	374	1687	2093

In this context, the well-known APTI database can be used only for Arabic text recognition in screen captures or in images extracted from PDF documents since it's made up of synthetic text images with a clean white background. In the same context, the recent database named Alif [1] is made up of artificial Arabic text images extracted from Arabic TV broadcast.

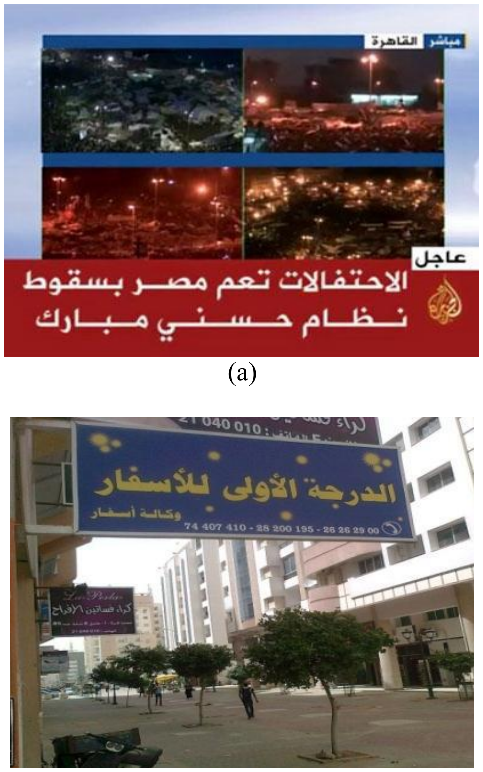
Contrary to this kind of artificial Arabic text, in which there is not a big variety in the emplacement of text, in text size, fonts and colors, in natural scene images, there are more strong difficulties in the recognition of text, due to the huge variance in size, font, color, non-planar surface, perspective distortion, *etc.*

Therefore, Alif database can't be useful in the case of Arabic text recognition in natural scenes images. Fig. 3 shows the challenges in recognizing text in natural scene images comparing in artificial ones.

III. Brief Description of Arabic Script Characteristics

Arabic character shapes depend on their position within a word (isolated beginning, middle, and end). Therefore, character models of the same character in different position ((fig4.a) and (fig4.b)) need to be discriminant from each other. Arabic script is rich and complex. Most notably:

- It consists of 28 letters written from right to left. It is cursive even when printed means that letters are connected.
- Arabic words consist of one or more sub-words called PAWs (Pieces of Arabic Word), PAWS without dots are called naked PAWS.
- It is cursive even when printed means that letters are connected.
- Some letters can be differentiated from each other's by dots like letters ’ح’ ,’خ’ and ‘ج’. Examples are shown in fig.4.
- In this case we say that these 3 letters form a joining group called Glyph. There are 19 Glyphs.
- Arabic characters depend on their position within a word (isolated, beginning, middle, and end). Examples are shown in (fig4. (1), fig4. (2), fig4. (3), fig4. (4)).



(a)

(b)

Fig 3. Challenges in recognizing text in (a) natural scene images comparing in (b) artificial ones (huge variance in size, font, color, non-planar surface , perspective distortion)

[1] https://cactus.orange-labs.fr/ALIF/index.html

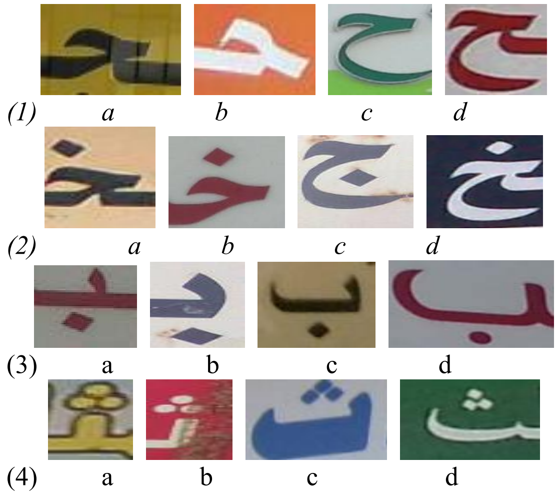
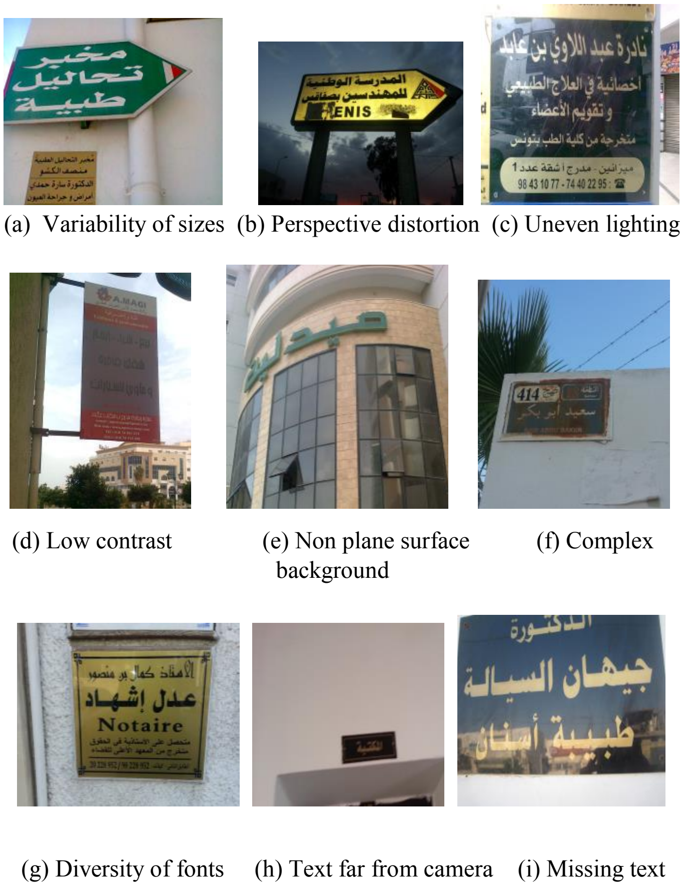


Fig. 4. Example of segmetend characters taken from ARASTI database

IV. DESCRIPTION OF THE DATASET

In this section, we present our database of Arabic scene text called ARASTI. In our work, we initially aimed to create an Arabic scene text database.



(a) Variability of sizes (b) Perspective distortion (c) Uneven lighting

(d) Low contrast (e) Non plane surface background (f) Complex

(g) Diversity of fonts (h) Text far from camera (i) Missing text

Fig. 5. Sample source images in our data set.

A. Arabic Text Image Dataset

Our database of the Arabic text is collected from pictures of shopping areas, signboards, roadside signs, *etc*.

We photographed a set of 374 images. They are taken under real life situations without caring about the environmental conditions.

Therefore, in images of our database, the lighting conditions may be more or less good, they can include blur, can be taken from different distances, *etc*. Some of these original images are shown in fig. 5.

B. Arabic Word Image dataset

The Arabic Word Image dataset contains 1280 cropped word images from Scene Texts. This dataset can be used for cropped word recognition. Some of these cropped words images are shown in fig. 6.

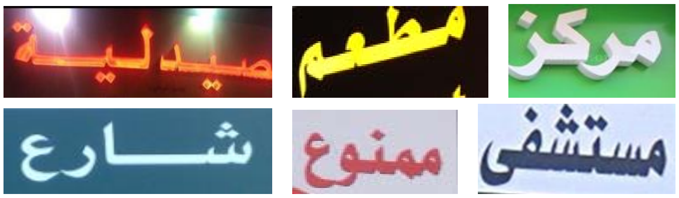


Fig. 6. Example of cropped Words taken from ARASTI database

C. Arabic Character Image dataset

To validate a character recognition method, we need to have a benchmark database covering all the varieties of characters. The different forms of these characters are used to store the character images in 55 classes. The different classes of characters are shown by fig 7. Therefore, we manually segmented individual characters to obtain a total of the 55 classes of the 28 Arabic characters in their different position within the word (start, middle, end and isolated) shown by table I. For each class we obtain between 20 and 57 samples.

The database contains an unequal number of samples for each characters due to the fact that in daily usage some characters are more frequently used than others. The category and number of samples collected there in are presented in Table I. ARASTI database is freely available and can be downloaded by consulting this web link: http://www.regim.org/publications/databases/arabic-scene-text-image-of-regim-lab2015/

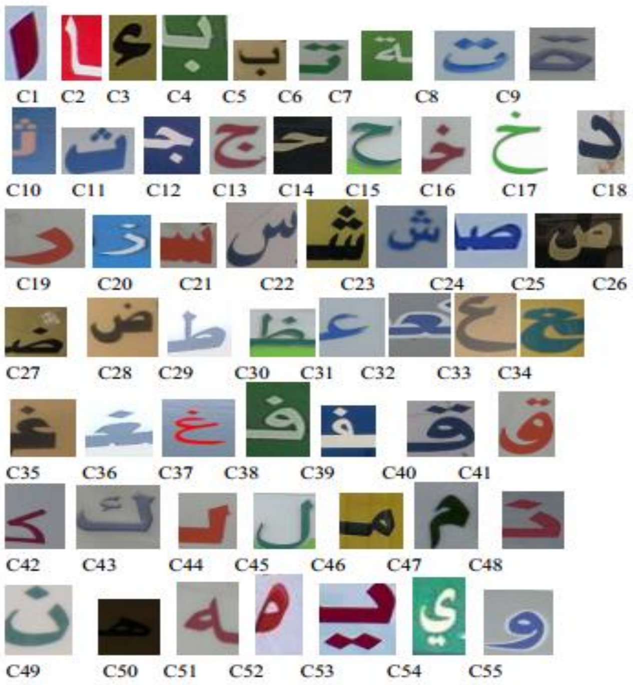

C1 C2 C3 C4 C5 C6 C7 C8 C9
C10 C11 C12 C13 C14 C15 C16 C17 C18
C19 C20 C21 C22 C23 C24 C25 C26
C27 C28 C29 C30 C31 C32 C33 C34
C35 C36 C37 C38 C39 C40 C41
C42 C43 C44 C45 C46 C47 C48
C49 C50 C51 C52 C53 C54 C55

Fig. 7 Examples of all character classes of ARASTI database

V. Conclusion

In this paper, a database of Arabic text captured from the wild is created. This database will serve as a benchmark for the future researchers in this direction. The ARASTI contains 374 images and 3740 scene character images. By the use of the dataset, differences among scene Arabic text recognition methods can be compared.

Using ARASTI database, we aim, in the future, to organize a competition to bring together the different researchers working on Recognition Arabic text in natural scenes in order to provide them a suitable benchmark to compare the performances of their different techniques.

Acknowledgment

This work is performed in the framework of a thesis MOBIDOC financed by the EU under the program PASRI. The authors would like also to acknowledge the partial financial support of this work by grants from General Direction of Scientific Research (DGRT), Tunisia, under the ARUB program.

References

[1] B. Epshtein, E. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” in Proc. of CVPR, 2010.

[2] L. Neumann and J. Matas, “A method for text localization and recognition in real-world images,” in Proc. of ACCV, 2010.

[3] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in Proc. of ICCV, 2011.

[4] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, “Detecting texts of arbitrary orientations in natural images,” in Proc. of CVPR, 2012.

[5] C. Yao, X. Bai, B. Shi, and W. Liu, “Strokelets: A learned multi-scale representation for scene text recognition,” in Proc. of CVPR, 2014.

[6] C. Yao, X. Bai, and W. Liu, “A unified framework for multi-oriented text detection and recognition,” IEEE Trans. Image Processing, vol. 23, no. 11, pp. 4737–4749, 2014.

[7] M. Jaderberg, A. Vedaldi, and A. Zisserman, “Deep features for text spotting,” in Proc. of ECCV, 2014. [8] Y. Zhu, C. Yao, and X. Bai, “Scene text detection and recognition: Recent advances and future trends,” Frontiers of Computer Science, 2015.

[8] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “ICDAR 2003 robust reading competitions,” in Proc. of ICDAR, 2003. [10] S. M. Lucas, “ICDAR 2005 text locating competition results,” in Proc. of ICDAR, 2005.

[9] A. Shahab, F. Shafait, and A. Dengel, “ICDAR 2011 robust reading competition challenge 2: Reading text in scene images,” in Proc. of ICDAR, 2011.

[10] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, “ICDAR 2013 robust reading competition,” in Proc. of ICDAR, 2013.

[11] T. E. de Campos, B. R. Babu, and M. Varma, “Character recognition in natural images,” in Proc. of VISAPP, 2009.

[12] A. Mishra, K. Alahari, and C. V. Jawahar, “Scene text recognition using higher order language priors,” in Proc. of BMVC, 2012.

[13] Digitally Represented Arabic Text," Document Analysis and Recognition (ICDAR), 2013 12th International Conference on , vol., no., pp.1433,1437, 25-28 Aug. 2013.

[14] A. Mishra, K. Alahari, and C. V. Jawahar, “Scene text recognition using higher order language priors,” in Proc. of BMVC, 2012

[15] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, C. L. Tan, “Multilingual scene character recognition with co-occurrence of histogram of oriented gradients”, Pattern Recogn. 51 (C) (2016) 125–134. doi:10.1016/j.patcog.2015.07.009.

[16] B. Zhang, W. Zhao, J. Liu, R. Wu, X. Tang, “Character recognition in natural scene images using local description”, in Intelligent Science and Intelligent Data Engineering- Second Sino-foreign-interchange Workshop, IScIDE 2011, Xi’an, China, October 23-25, 2011, Revised Selected Papers, 2011, pp. 193–200.