SmartATID: A mobile captured Arabic Text Images Dataset for multi-purpose recognition tasks

Fatma Chabchoub, Yousri Kessentini, Slim Kanoun, Véronique Eglin

SmartATID: A mobile captured Arabic Text Images Dataset for multi-purpose recognition tasks

Fatma Chabchoub, Yousri Kessentini[1,2], Slim Kanoun and Veronique Eglin[3]
1 MIRACL laboratory, ISIMS, university of Sfax, Tunisia
2 Digital Research Center of Sfax, Tunisia
3 LIRIS laboratory, INSA de Lyon, France
{chabchoub.fatma, yousri.kessentini, slim.kanoun}@gmail.com
veronique.eglin@insa-lyon.fr

Abstract —**Today's smartphones are able to capture documents with a good and simple way as any personal scanners. The captured document images need to be processed by specific and automated document processing systems. The systems are dedicated to textual content analysis, indexing and recognition. For instance, they may be used for font identification, writer identification and word or line segmentation. The state-of-the-art works lack comprehensive database for Arabic document images which are captured by mobile phones. This paper presents the first public offline images database for both printed and handwriting Arabic mobile captured documents, named "SmartATID". The document images of the database are acquired under varying capture conditions (blur, perspective angles and light). This causes photometric and geometric distortions that influence the performance of OCR process but also the page segmentation in lines and paragraphs. Each document image of our database is provided with a ground truth file that contains the exact text transcription and all numerical capture parameters used for each image capture. The database is freely and publicly usable by the research community at the following address http:// sites.google.com/site/smartatid.**

Keywords-Smartphone Arabic document capture database; mobile OCR; capture-based distortions

I. INTRODUCTION AND RELATED WORK

Document image processing is today considerably influenced by evolutions of data acquisition systems and embedded technologies for new innovative mobile applications. Especially, the smartphone revolution has impacted the way people use information and at the very beginning of the processing flow the way they digitize them. A smartphone-captured document image is at the origin of the new challenges associated to content image analysis and recognition by automated embedded document image processing systems ([1], [2] and [3]). These systems need to deal with the camera sensors, the resolution of the images, the light conditions, the physical distortions of the paper (folds etc.), the perspective distortion, the out-of-focus blur and/or the motion blur during capture [1].

In order to face this new challenge, we all need to develop reliable solutions for improving best suited image quality required to efficient embedded text recognition systems. Recognizing texts from images captured by mobile phones offers access to significant data. By using those data, we intend to perform new information retrieval system, reuse contents on demand and share them through networks and users. In that context, it becomes a necessity to make available standard databases that closely match the real world scenarios and offers new services to testify the new embedded mobile systems. Such standardized databases serve to help researchers to develop their solutions. Besides, they provide a baseline benchmark between different systems within evaluation campaigns and competitions.

The analysis of the literature shows an absolutely terrible lack of databases and baseline benchmarks for digitizing mobile captured documents. Indeed, the only available dataset comprising of documents captured in real conditions with various photometric and geometric distortions was generated for ICDAR2015 Competition on Smartphone Document Capture and OCR by Burie et al. [4]. This dataset contains 12100 document images captured from 50 different paper documents with real content from wiki-books and cooking recipes from Internet. All documents contain single column English text printed with multiple scales, fonts, font-faces and colors. At least 240 different images are taken per document under varying capture conditions (perspective angles, light and blur).

Special mention may be made for the research community working on quality assessment of smartphone-captured-document images who created recently two datasets for evaluating and benchmarking their methods.

The first one was proposed in 2013 by Kumar et al. [5]. This dataset deals with only one type of documents, one type of capture distortions (out-of-focus blur) and the images are captured using one smartphone. The dataset of Kumar et al. [3] has 29 different documents used to capture 375 images with varying degrees of out-of-focus blur.

The second one was named Smart Doc-QA and was created in 2015 by Nayef et al. [6]. This dataset deals with

different real paper document types, single and multiple capture distortions and the images are captured using two smartphones. Smart Doc-QA [6] contain 4260 document images captured from 30 different paper documents with representative value of different distortions (Light, perspective and blur).

A deep observation on the state-of-the-art of existing datasets for mobile captured documents shows there is no standard Arabic datasets for both printed and manuscript contents. Consequently, this paper presents a detailed description of the constitution of an Arabic multi-font and multi-size printed/handwritten text database dedicated to recognition and content based information retrieval tasks.

Our mobile printed and handwritten Arabic document images database is a collection of smartphone-captured documents. This database will be useful mainly for document analysis scenarios (page and line segmentation tasks), Arabic text recognition and also could be helpful for assessing document image quality, and improving it.

The paper is organized as follows: in section 2, we present the overall acquisition protocol and the different distortions correlated to a mobile images capture. Then in section 3 we present the ground truth representation including the text transcription encoding and different tags for the acquisition parameters required for further recognition tasks evaluations.

II. OVERVIEW OF "SMARTATID" DATABASE

In this section, we present our database of Arabic printed/handwritten texts for mobile captured documents.

A. Documents image selection

In order to create the Smartphone-captured Arabic printed text images dataset, we selected a set of document images from APTID/MF database [7] (Arabic Printed Text Image Database /Multi-Font). This selection is composed by 32 text-blocks images which are scanned at 300 dpi resolution in grayscale format and whose content is organized in many sizes and fonts among the 10 fonts presented in Fig. 1. The set of selected images are printed with a Brother printer, in total there are 116 paper documents with multiple scales and fonts used to capture 16472 images with mobile phones for our printed text image Dataset. Fig. 2 shows examples of those paper documents.

With the same process applied to create the dataset of Smartphone-captured Arabic printed text images, we also build the Smartphone-captured Arabic handwritten text images dataset. The Arabic handwritten documents are selected from ALTID database [8] (Arabic/Latin Text Images Database for recognition research). As the handwritten dataset of ALTID database is inherited from the printed document pages of APTID/MF database, we choose to use the same handwritten text images than the Smartphone-captured Arabic printed text ones. This selection is made up of four Arabic handwritten texts written by eight writers and presented according two layouts:

- a formal layout: the author writes the text in a formal style using a lined paper as a background template with 2cm for line spacing.
- an informal layout: the author rewrites the same text in a freely without any layout constraints.

Consequently, there are 64 documents (32 different formal documents and 32 different informal documents) used to capture 9088 images with mobile phones of our handwritten text image Dataset. Fig. 3 shows examples of those handwriting documents presented with layout constraints and without any.

Figure 1. 10 Arabic fonts used in APTID/MF [5].

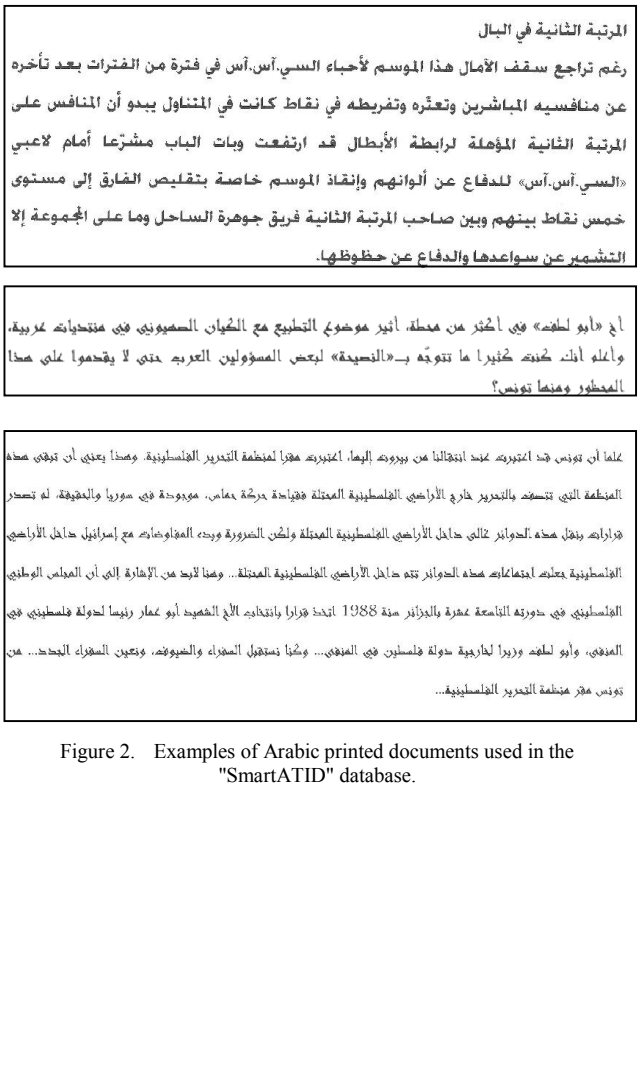


Figure 2. Examples of Arabic printed documents used in the "SmartATID" database.

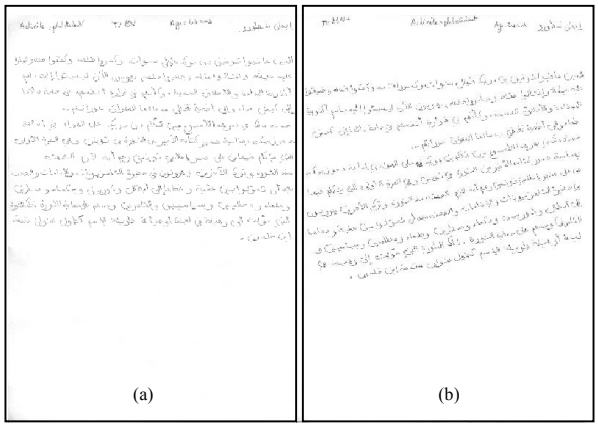


Figure 3. Examples of Arabic handwritten documents used in the "SmartATID" database: (a) Formal (constrained) handwritten text, (b) Informal (unconstrained) handwritten text.

For our two datasets, the documents have been chosen with plain and simple background in order to reduce the impact of the background effects on the automatic pages processing (i.e.: OCR results). Anyway the quality of the captured document images has a big influence on OCR performance.

B. Logistics for the data capture

Despite the lack of automatic material for the acquisition, we succeeded in capturing the document images for the database in a precisely repeatable and controlled environment. As a consequence, the majority of steps have been performed manually in real conditions by setting various photometric and geometric distortions to images.

In our dataset we mainly examine scene-related distortions such as light conditions, out-of-focus blur, motion blur and perspective angles distortions. To ensure real life scenarios of the possible light conditions, we performed image capture outdoors and image capture indoors with strictly controlled light conditions. In order to simulate the possible capture positions, the out-of-focus blur and the motion blur during the captures, we have employed a rolling adjustable height and swinging Camera.

For the device-related distortions, we have employed two modern smartphones, Samsung Galaxy S6 edge and I phone 6S plus, whose cameras are fitted with different sensor technologies and able to capture images at different resolutions (13MP and 8MP).

The sequence of steps dedicated to the capture is synchronized by a human so as to ensure the generation of digitized data in real situation. As examples of human interactions, we can mention: the manual change of paper document, of light conditions and the location of the focal distance from the smartphone camera and the center of the document page. Once a desired position and other capture conditions are achieved, the agent responsible to the image capture communicates with the smartphone through a Bluetooth to trigger the capture.

Figure 4. The rolling support holding smartphone over a document on a fixed simple background.

Fig. 4 presents a photo of the rolling support holding the smartphone beside a document ready to be captured.

C. Capture Protocol

As mentioned previously, our Arabic database contains two datasets: one is composed by smartphone-captured Arabic printed text images and the second by smartphone-captured Arabic handwritten text images. The first smartphone-captured Arabic printed document dataset has a total of 16472 document images captured from 116 different paper documents. Hundred and forty two different images are captured per document (71 captures per phone). The second smartphone-captured Arabic handwritten document dataset has a total of 9088 document images captured from 64 different paper documents. In the same way than for the printed dataset, 142 different images are captured per document (71 captures per phone).

During the creation of both datasets, the captures are taken with representative values of different types of distortions (see capture parameters below). For each image, the information about the document content and the capture conditions are stored as ground truth for upcoming evaluation.

1) Fixed Capture Parameters:

The documents capture is established through different fixed parameters setting that guarantee a controlled acquisition and similar for all images of the dataset.

- Background: simple and plain with a contrasted color with the paper documents to be captured. The background is chosen to facilitate the page border detection process.
- Document layout and dimensions: A4 paper size with a fixed orientation.
- Smartphone setting: flash is always deactivated

Fig.4 shows the fixed capture parameters related to the documents and the background.

2) Variable Capture Parameters:

The documents capture is also established through different variable parameters setting that ensure multiple acquisition conditions and offer the desired diversity of images for the dataset. The detailed capture protocol and integration of these constraints are presented in section 3 and 4.

- Smartphone camera: 2 smartphones
- Light: 5 light conditions
- Out-of-focus blur: 4 values
- Motion blur: 2 types. We use two motions blur, a horizontal one and a 2D one.
- Position of the smartphone camera with respect to the paper document (5 positions):
  - o Perspective 1: Longitudinal incidence angle (mobile rotation around Y-axis).
  - o Perspective 2: Lateral incidence angle (mobile rotation around X-axis).
  - o Distance between the camera and the document: 1 value (35cm)

In the following section, we detail our protocol for images capture and we detail the choice of specific values for the variable capture settings mentioned above. The captures are organized into two categories: single captured distortions (singly distorted document images) and multiple captured distortions (multiply distorted document images). The presence of multiple causes of distortions is undoubtedly the expression of the reality of the ground. Nevertheless, possessing singly distorted document images is useful to assess the impact of each type of distortions taken individually for the evaluation of OCR results.

For both distortion categories, we take every capture with the following two smartphones:

- Samsung Galaxy S6 edge (camera: 13MP)
- IPhone 6S plus (camera: 8MP)

3) Single Distortions:

Under “single distortions” constraints, we study independently light conditions, out-of-focus blur and motion blur. All images with single captured distortions are captured at the same position, called "parallel position with a minimal perspective distortion” (i.e.: a position where the perpendicular distance between the center of the document and the camera of the smartphone is 35cm and both angle of mobile rotation around Y-axis and of mobile rotation around X-axis equal zero. Under these conditions, we perform the capture process as follows:

For light distorted images, five successive images are taken for each document under five light conditions mentioned below.

- Light condition 1: Day light only (without any artificial lights) outdoors.
- Light condition 2: Day light only (without any artificial lights) indoors.
- Light condition 3: Day light + Economic lamp Led indoors.
- Light condition 4: Night + table lamp light.
- Light condition 5: Table lamp light + an object casting a shadow on a large part of the document.

For out-of-focus blur, all the images are also captured at the same parallel position presented above, with the light condition fixed to "day light + Economic lamp Led indoors". The smartphone camera is focused on the document at a distance closer than 35-cm (22-cm), then the image is captured at the position with a 35-cm distance. This leads to a controlled out-of-blur capture of the pages. The same operation is performed four times in order to capture four different images at varying out-of-focus blur degrees. Each time, the focus point is lcm shorter to the document, whereas the capture is taken at the same distance of 35cm.

For motion blur, we consider the same light condition and camera position as the out-of-focus blur captures. The motion blur is executed as follows. While the rolling support is moved at a specific speed and aimed to a certain direction, the smartphone camera is started to take the capture. The capture can occur at any moment during the motion, hence producing images with different degrees of severity of blur. Two images are captured this way for each document (image with horizontal motion blur and image with 2D motion blur).

Fig. 5 shows three different images of the same document captured at the third light condition and the parallel position. The first image presents a focused (sharp) capture, the second presents a blurry image due to out-of-focus blur and the third presents a blurry image due to motion blur. Finally the overall process has produced a total of 11 captures. The same operation is repeated for each smartphone camera, where we ensure that each capture shows only one type of distortion.

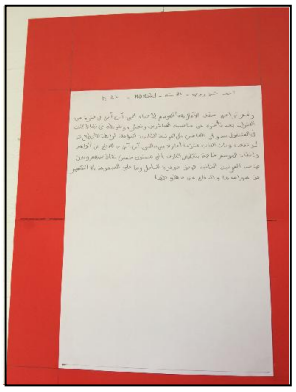
4) Multiple Distortions:

Here we use a combination of different capture conditions for each image. We have chosen 3 light conditions, 5 camera positions and 3 blur values. Additionally, we take a reference capture, which is not blurry focused or blurry sharp, at each combination of light condition and position. This creates 60 captures per document for each dataset.

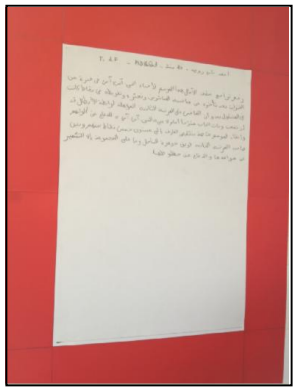
(a) A sharp image-Focused. (b) A blurry image-out-of-focus. (c) A blurry image-2D motion.

(d) Cropped image (a). (e) Cropped image (b). (f) Cropped image (c).

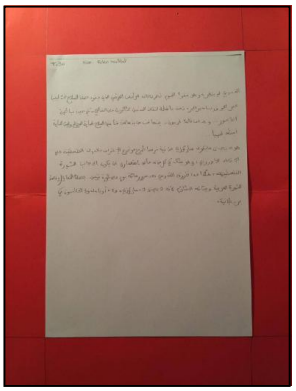
Figure 5. Examples of captured images for single distortions.



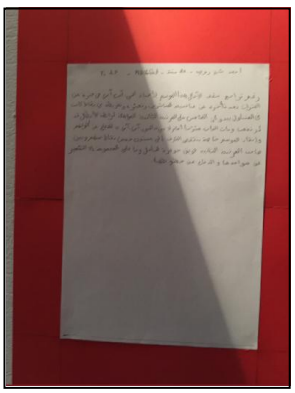
(a) Capture position (5°,10°,35cm) light condition #2, sharp (focused)



(b) Capture position (-5°,10°,35cm) light condition #3, out-of-focus blur



(c) Capture position (0°,0°,35cm) light condition #4, sharp (focused)



(c) Capture position (-10°,5°,35cm) light condition #5, 2D Motion blur.

Figure 6. Examples of captured images for multiple distortions.

For the light conditions, we have selected conditions 3, 4 and 5 mentioned section 3. We have chosen to express the position of the camera with respect to the document in the following way (perspective 1, perspective 2, and distance), in respect to what has been mentioned above in the variable capture parameters. We have also considered five different positions: (0°, 0°, and 35cm), (_10°, _5°, 35cm), (_10°, 5°, 35cm), (_5°, 10°, 35cm) and (5°, 10°, 35cm) combined with only two of the out-of-focus blur values mentioned above and the 2D motion blur. Fig. 6 shows examples of captured images for multiple distortions.

D. Ground Truth File drscription

An essential component of any database is the presence of ground truth data [9].

Each image of our two datasets, the smartphone-captured Arabic printed documents dataset and the smartphone-captured Arabic handwritten documents dataset, is provided with the following ground-truth information:

- The transcription of the text in a document.
- The capture parameters (distortion types and values) and the ID of a captured document.
- A sharp "reference" image of the document at each combination of position and light.

An example of such XML file for our smartphone-captured Arabic handwritten text images dataset is given in figure Fig. 7.

Figure 7. Example of XML file present the ground truth information of smartphone-captured Arabic handwritten text images dataset.

The XML file is composed by three principal markups sections:

- Content: in this element, we have the document write type (Formal layout or Informal layout), the number of line of Arabic text (nbline) and sub-elements for each line where we have her identical transcription and the number of her content tokens.

- Smartphone: in this element, we specify the type of camera and the resolution value used to capture the image document.

- Distortion: in this element, we indicate the category of the capture (singly distorted document images or multiply-distorted document images) and we present a sub-elements for each capture parameters such as: light condition, out-of-focus blur, motion blur and perspective.

Each text image in our smartphone-captured Arabic printed text images dataset is fully described using the same Xml file with a simple difference at the "Content" element where we specify the font name, the number of line of Arabic text (nbline) and sub-elements for each line where we have her identical transcription and the number of her content tokens.

III. CONCLUSION

In this paper we address the lack of Arabic databases for mobile captured documents, we have created a novel database, named Smart ATID, which contains an off-line Smartphone-captured Arabic printed text images dataset and an off-line Smartphone-captured Arabic handwritten text images dataset. This database proposes both single and multiple capture distortions for subsets of the images, allowing to address the specific issue of real conditions. Additionally, it includes a complete ground-truth with type and amount of each distortion contained in the images and reference images. We hope this database will serve the scientific community for many years to come for benchmarking mobile Arabic recognition methods.

Our future work consists in the improvement of human motion reproduction and the extension of our two datasets to more document types, more smartphones and more distortions.

REFERENCES

[1] J. Liang, D. Doermann, and H. Li. "Camera-based analysis of text and documents: a survey" International Journal on Document Analysis and Recognition, 7(2-3), 2005, pp. 84-104.

[2] F. Shafait and T. M. Breuel, “Document image dewarping contest,” in 2nd Int. Workshop on Camera-Based Document Analysis and Recognition (CBDAR), 2007, pp. 181–188.

[3] J. Liang, D. DeMenthon, and D. Doermann, “Geometric rectification of camera-captured document images,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 30, no. 4, pp. 591–605, 2008.

[4] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. Luqman, M. Mehri, N. Nayef, J. Ogier, S. Prum, and M. Rusinol, "lcdar2015 competition on smartphone document capture and ocr (smartdoc)," in ICDAR, 2015, pp. 1161-1165.

[5] J. Kumar, P. Ye, and D. Doermann, "A dataset for quality assessment of camera captured document images," in CBDAR, 2014, pp. 113-125.

[6] N. Nayef, M. Luqman, S. Prum, S. Eskenazi, J. Chazalon and J. Ogier, "SmartDoc-QA: A Dataset for Quality Assessment of Smartphone Captured Document Images - Single and Multiple Distortions", in CBDAR, 2015, pp. 1231-1235.

[7] F. Jaiem, S. Kanoun, M. Khemakhem, H. El Abed, and J. kardoun, "Database for Arabic Printed Text Recognition Research," ICIAP 2013, Part I, LNCS 8156, pp. 251-259,2013.

[8] I. Chtourou, A. Cheikh Rouhou, F. Jaiem, and S. Kanoun, "ALTID: Arabic/Latin Text Images Database for recognition research", in ICDAR, 2015, pp. 836-840.

[9] S. Ahmed, M. Imran Malik, M. Zeshan Afzal, K. Kise, M. Iwamura, A. Dengel and M. Liwicki, "A Generic Method for Automatic Ground Truth Generation of Camera-captured Documents", in arxiv.org, 2016.