World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:13, No:4, 2019

Data Gathering and Analysis for Arabic Historical Documents

Ali Dulla

***Abstract*—**This paper introduces a new dataset (and the methodology used to generate it) based on a wide range of historical Arabic documents containing clean data simple and homogeneous-page layouts. The experiments are implemented on printed and handwritten documents obtained respectively from some important libraries such as Qatar Digital Library, the British Library and the Library of Congress.We have gathered and commented on 150 archival document images from different locations and time periods. It is based on different documents from the 17th-19th century. The dataset comprises differing page layouts and degradations that challenge text line segmentation methods. Ground truth is produced using the Aletheia tool by PRImA and stored in an XML representation, in the PAGE (Page Analysis and Ground truth Elements) format. The dataset presented will be easily available to researchers world-wide for research into the obstacles facing various historical Arabic documents such as geometric correction of historical Arabic documents.

***Keywords*—**Dataset production, ground truth production, historical documents, arbitrary warping, geometric correction**.**

I. Introduction

ALONG with this growth in digital information, the demand for digitising historical paper documents has become essential for most libraries and museums; however; there is increasing anxiety over the performance of OCR systems, which depends heavily on the quality of the document images that are frequently influenced by various geometrical distortions [1]. The utility of digitized documents is directly related to the quality of the extracted text. Principle archives of historical documents are being gathered around the world. This requires an exact translation of the contents for automation of engaging metadata, full-text searching, and data extraction. [2]. As a result of this digitisation, unwanted distortion may be present in the final images, such as page curl, arbitrary warping (see Fig. 3), and folds. Page curl: The curvature of the pages towards the spine is referred to as page curl. Arbitrary warping: Wavy curves in document image are referred to as arbitrary warping. Fold: The bending of a document image is referred to as fold. All these deformations will be problematic for commercial OCRs that are tailored for flat pages with straight text lines. Also, modest warping can cause most current OCR systems to fail [3]. The most popular Arabic dataset is the IFN/ENIT database [4]. It involves of cities and villages names written by 411 writers, with 946 different names, totaling 26,459.handwritten names containing more than 210,000 characters. There are also the APTI and KHATT databases. APTI for Arabic Printed Text Images [5]. The database comprises 45'313'600 single word Images totaling to more than 250 million characters. Handwritten Text database (KHATT) [6]. The database holds the group of 1000 handwritten forms written by 1000 writers from different countries. On the other hand there is no dataset of warped historical Arabic documents.

A. History and Development of the Arabic Script

Generally, there are two schools of thought regarding the origin of the present Arabic script. One believes it is Nabataean (of Petra in Jordan), whereas the opposite insists it's Syriac. Before the appearance of Islam, Arabic writing was in use within the sixth century, within the Arab kingdom of the Syria-Mesopotamia region (mainly in Al-Hira) yet as in Mecca. These sources recommend that the Arab writing system was "derived from the Syriac and … had originated in Iraq or Midian [7]. Arab historians believe that the primary step within the development of Arabic hand was ancient Egyptian hand (see Figure 1). Then, Finiqi was generated from Egyptian [7]. Arabic writing was without dots and diacritics. When the Holy Quran was written, in Arabic, for the first time, it was without dots and diacritics. Also, it was allowed to continue the same word on the following line. This practice is allowed in modern Latin languages but not in current Arabic writing [8]. Various Arabic calligraphic styles developed in varied Arabian cities, with completely different writing techniques and writing tools. The most known Arabic calligraphic styles see Fig. 2.

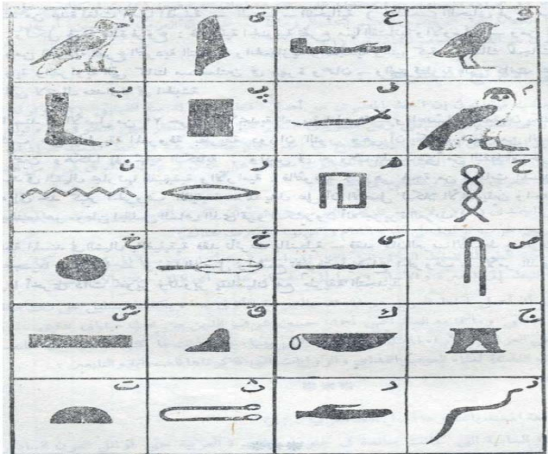


Fig. 1 Alphabet of ancient Egyptian penmanship

A. Dulla. is with the School of Computing, Science and Engineering, University of Salford, Greater Manchester, United Kingdom (corresponding author, phone: +447448740489; e-mail: A.Dulla@edu.salford.ac.uk).

Fig. 2 The most Scripts of Arabic Calligraphy

(a)

(b)

Fig. 3 (a) The original image (b) The warped image (images with page curl)

B. Comparing Arabic with Latin Font Features

According to [9], Arabic is now spoken by almost than 422 million people across the world and it is understood as a cultural emblem for a number of people. However, in comparison to the printed Latin script, Arabic has certain variations, as the letters change according to their position in the word. In fig.4, the different shapes of 28 Arabic letters are provided in accordance with their word positions. Latin and Arabic writings are the most popular scripts used around the world. This part highlights the differences and similarities between the Latin and Arabic writings as well as their typographical features [10]. We shall discuss several points which can be studied in the design process.

- Direction of writing

The writing direction in Arabic starts from the right side toward the left side whereas numerals are written from left to right [10]. This differs from Latin text, where the writing direction goes from left to right also for the numbers.

- Connectivity and cursiveness

The characters of Latin and Arabic writings are used differently to form words. In Latin writing, characters can be used as independent letters or they can be joined in cursive style. The cursive style may be used just for authenticity or imitating handwriting form. Unlike Arabic, there is no difference between printing and handwriting and only the cursive style is accepted. Fig. 5 enumerates all characters and their shapes in isolated form and in their initial, middle and final forms respectively [10].

- Ligatures

Ligature happens where two or more characters are united as a single glyph. In Latin, some ligatures are stylistically used when consecutive characters collide with each other. In Arabic writing, using ligatures is very widespread. One of the most used ligatures is ﻻ.

- Diacritical marks

Arabic writing is very wealthy in diacritic symbols. The presence or the absence of these diacritics distinguishes the same

main shape of Arabic letters. In Latin, they are used in all Latin languages excluding English. They are used in languages whose phonetics use sounds that do not exist in Latin [9].

No	Letter label	Isolated	Begin	Middle	End
1	Alif	ا		ا	
2	Baa	ب	بـ	ـبـ	ـب
3	Taaa	ت	تـ	ـتـ	ـت
4	Thaa	ث	ثـ	ـثـ	ـث
5	Jiim	ج	جـ	ـجـ	ـج
6	Haaa	ح	حـ	ـحـ	ـح
7	Xaa	خ	خـ	ـخـ	ـخ
8	Daal	د		د	
9	Thaal	ذ		ذ	
10	Raa	ر		ر	
11	Zaay	ز		ز	
12	Siin	س	سـ	ـسـ	ـس
13	Shiin	ش	شـ	ـشـ	ـش
14	Saad	ص	صـ	ـصـ	ـص
15	Daad	ض	ضـ	ـضـ	ـض
16	Thaaa	ط	طـ	ـطـ	ـط
17	Taa	ظ	ظـ	ـظـ	ـظ
18	Ayn	ع	عـ	ـعـ	ـع
19	Ghayn	غ	غـ	ـغـ	ـغ
20	Faa	ف	فـ	ـفـ	ـف
21	Gaaf	ق	قـ	ـقـ	ـق
22	Kaaf	ك	كـ	ـكـ	ـك
23	Laam	ل	لـ	ـلـ	ـل
24	Miim	م	مـ	ـمـ	ـم
25	Nuun	ن	نـ	ـنـ	ـن
26	Haa	ه	هـ	ـهـ	ـه
27	Waaw	و		و	
28	Yaa	ي	يـ	ـيـ	ـي

Fig. 4 Arabic letters

Letter Name	Transliteration	Pronunciation & English Equivalent	Contextual forms			
			final	Middle	Initial	Isolated
alif	*A*	-	ـا	-	-	ا
baa	*B*	B as in bake	ـب	ـبـ	بـ	ب
taa	*T*	T as in take	ـت	ـتـ	تـ	ت
thaa	*Ṯ*	th as in thin	ـث	ـثـ	ثـ	ث
jiim	*Ǧ*	j as in joke	ـج	ـجـ	جـ	ج
Haa	*ḥ*	no equivalent	ـح	ـحـ	حـ	ح
khaa	*ḫ* (also *kh, x*)	no equivalent	ـخ	ـخـ	خـ	خ
daal	*d*	d as in day	ـد	-	-	د
dhaal	*ḏ* (also *dh, ð*)	th as in this	ـذ	-	-	ذ
raa	*r*	r as in car	ـر	-	-	ر
zaay	*z*	z as in zeal	ـز	-	-	ز
siin	*s*	s as in snake	ـس	ـسـ	سـ	س
shiin	*š* (also *sh*)	sh as in shake	ـش	ـشـ	شـ	ش
Saad	*ṣ*	emphatic s	ـص	ـصـ	صـ	ص
Daad	*ḍ*	emphatic d	ـض	ـضـ	ضـ	ض
Taa	*ṭ*	emphatic t	ـط	ـطـ	طـ	ط
DHaa	*ẓ*	emphatic dh	ـظ	ـظـ	ظـ	ظ
ayn	ʿ	no equivalent	ـع	ـعـ	عـ	ع
ghayn	*ġ* (also *gh*)	no equivalent	ـغ	ـغـ	غـ	غ
faa	*f*	f As in face	ـف	ـفـ	فـ	ف
gaaf	*q*	emphatic k	ـق	ـقـ	قـ	ق
kaaf	*k*	k as in key	ـك	ـكـ	كـ	ك
laam	*l*	l as in leaf	ـل	ـلـ	لـ	ل
miim	*m*	m as in make	ـم	ـمـ	مـ	م
nuun	*n*	n as in none	ـن	ـنـ	نـ	ن
haa	*h*	h as in hat	ـه	ـهـ	هـ	ه
waaw	*w* / *ū* / *aw*	w as in wake	ـو	-	-	و
yaa	*y* / *ī* / *ay*	y as in yell	ـي	ـيـ	يـ	ي

Fig. 5 Different forms of Arabic letters

II. Creating the Dataset

We intended to create a database of historical Arabic script which would include unwanted distortion that may be present in the final images, such as page curl and arbitrary warping. Dataset constructing is based on two core actions that were chosen and the delivery of images, as well as metadata and dataset description, supported by ground truth creation for distinct subsets.

A. Image Gathering

The image Gathering process for producing a dataset of such a size had to be very directly and strictly defined and followed. To make sure that the dataset comprises a sufficient number of documents with both simple and complex layouts in each of the content categories, the first step of document choice is an off-line expert-driven activity. We carried out this step by following strict protocols. First, the collected documents which had been scanned by libraries at 350 dpi and in 24-bit colour have been processed later. The second step image choice was accompanied by metadata collection (see below). Owing to privacy issues, it is not easy to gain access to a large number of printed Arabic historical documents. Also, this is due to the lack of many kinds of distortion of historical Arabic documents. Subsequently, reasonable care is taken to create each document synthetic images. They are generated from a few base images by warping text lines close to the ones above, and text lines, as well as overlap area, are finely ground truths. To measure whether the stages of proposed system have any harmful effect images with straight lines were selected. Once every accessible data was of satisfying quality, the pre-processing of the images could be carried out in order to be ingested into the dataset. The source images were presented in differing formats, such as uncompressed TIFF and JPEG.

It was, therefore, important to standardise to an open and easy-to-use format so that images in the dataset could be easily used by a wide range of tools.

B. Ground Truth Creation

One of the important advantages of this dataset is the important amount of detailed and high quality ground truth available and the scope for its use. Ground truth, in this context, is a particular and formalized replica of what is really present on the physical page or, to place it in different words, what the right analysis/recognition approach is expected to come as result. Accordingly, it has to be designed (or at least tested) by a human. Aletheia is a sophisticated system for precise and yet practical ground truthing of a lot of documents. It bolsters top-down and base up ground truthing [11]. The format of the ground truth changes depending on the responsibility it is connected to (such as region outlines for segmentation or Unicode text for OCR). The ground truth is saved in the XML construction which is a section of the PAGE (Page Analysis and Ground truth Elements) description structure [12]. Ground truth is a vital advantage not only for improving and training new methods (such as adaptive text line segmentation or lexicon-based post correction) but also for performance evaluation as well as modification of end-to-end digitization workflows. We tested the evaluation method on different manually created grids for a warped historical document to prepare for the different percentage of warping on the image such as 0% warping, 5% warping, 25% warping, 50% warping and 75% warping. Ground truth grids and automatically generated dewarping grids are matched line by line. Each horizontal line of one set of grids is therefore matched against a nearby line of the other set of grids as shown in Fig. 6.

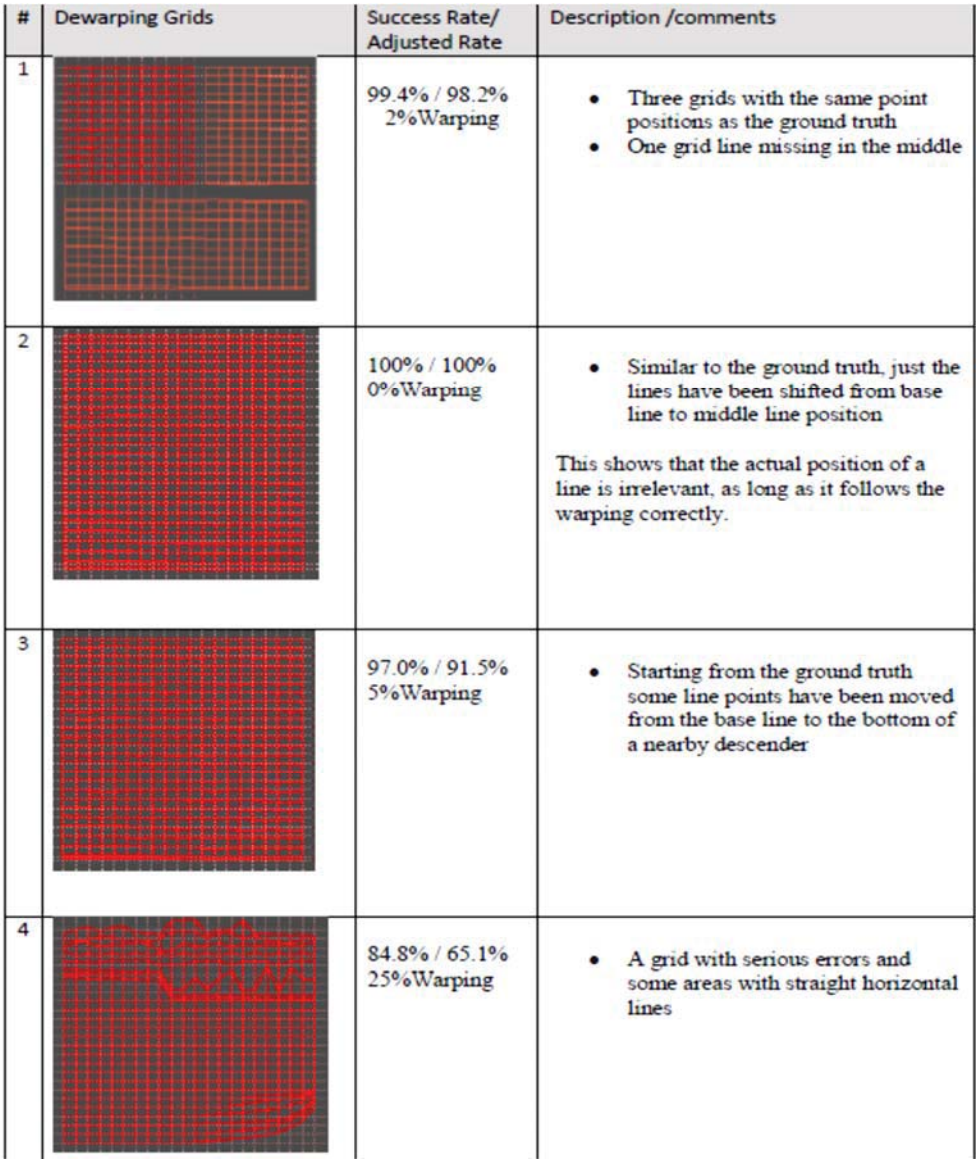


Fig. 6 Generated dewarping grids are matched line by line

Utilizing the ground truth baseline as reference, two error values are calculated, based on Average distance to result baselin and Average slope (angle) difference. The two values are calculated for each horizontal pixel position of a text line and combined using the arithmetic mean. The success rate for one single baseline is the harmonic mean of the non-linear distance success and the angle difference success.

III. Dataset Description

The dataset contains a wide variety of documents reflecting both the holding of major Arabic and some European libraries. The dataset includes over 200 images originating from four different national and significant libraries across the world. such as Qatar Digital Library, the British library, the Library of Congress and Juma Al-Majid Center for Culture and Heritage. A list of all the content providers forward including numbers of contributed images is presented in Table I.

TABLE I
A List of Including Numbers of Contributed Images

Library	Country	Number
Qatar Digital Library	Qatar	100
British library	UK	50
Library of Congress	USA	40
Juma Al-Majid Center for Culture and Heritage	UAE	10

A. Images

The dataset covers variable page layouts and degradations that challenge text line segmentation and baseline detection techniques. Disregarding the page layout typically leads to an under-segmentation of text lines. The erraticism of page layouts in historical documents is much bigger than in the current book. In addition to some forthright features such as headings, subtitles, paragraph divisions, page numbers and page separators, there are some specific features such as ornamental letters, drop letters and catchwords. In terms of age, the majority

of the manuscripts in the dataset –almost 90% – were produced in the 19th or early 20th century (particularly in the case of newspapers). Moreover, 23% were produced in the 17th or 18th century, with the remainder of the images ranging as far back as the 15th century. Although all probable steps were used to avoid some issues, in a dataset it is inevitable. There is a very small percentage of images where the publication year is not known, or was not made available to us (marked as "?" in the table below). A complete itemization of the age of the original documents in combination with the document types is presented in Table II.

TABLE II
DOCUMENT TYPE AND PRODUCTION CENTURY DISTRIBUTION

Century	Book Page	Newspaper page	Legal Document Page	Journal Page	Other Document Page	Unclassified Page	TOTAL
17	10	0	0	0	2	0	12
18	20	5	0	0	0	0	25
19	30	5	0	0	0	3	38
20	60	10	10	5	5	0	90
?	25	5	0	0	4	1	35

B. Metadata

The metadata is an XML-based representation that represents information at the document, page, and zone levels. It has many uses including merging/splitting functions and reading order.

In order to allow users to professionally search the repository, a huge set of metadata is preserved as part of the dataset. All accessible metadata is indexed and can be used as search parameters to access particular images or sets of images within the total dataset. The metadata has been designed under the following system:

- Digitisation data — resolution, bit depth, image dimensions, scanner used, file type, compression algorithm and quality, source of digitisation (paper, microfilm, etc.).
- Bibliographic knowledge — title, author, publicatiodate and location, document type, page number,
- Physical properties — language, script, typeface and number of columns.
- Copyright data — copyright holder, contactdetails, publishing permissions, content provider's reference.
- Managerial information original filename, access log,

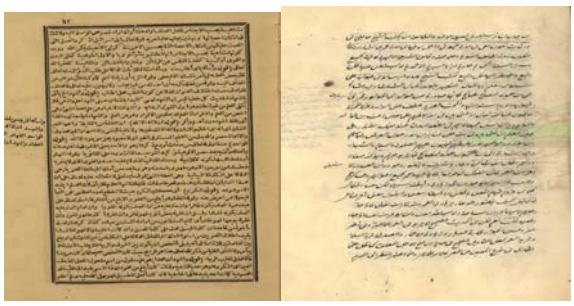
In addition to image-linked metadata, part of the dataset has a number of extra keywords related with each image. That list of keywords was collected with input from different libraries in a way that it provides additional useful search possibilities for different users of the dataset.

- Condition related — stains, holes, missing parts, tearsfolds, etc.
- Document related — impressions, filled in characters, broken characters, blurred, faded, etc.
- Scanning related — skew, parts of adjoining page, fingers, paper clips, copyright notices, etc.

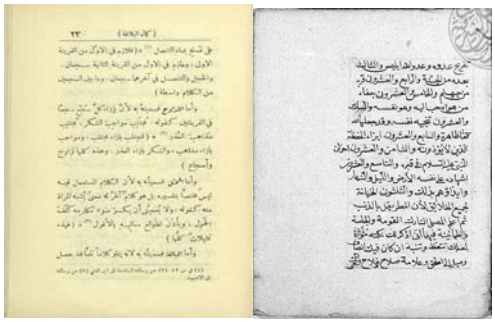
C. Grid Characteristics

Grid-based model, on this topic, can be used to carry out warping. The grid displays the warped surface of the document image and is applied for the geometric correction. It is defining as a matrix of two-dimensional points grid cells. For the de-warping, the regions of the image corresponding to cells are transformed to a rectangle. Given the grid is incorrect, this results in a warped document image. For extra adaptability, the data structure provides various grids for one document, with the restriction that the bounding boxes of the grids do not overlap. The grid data structure can be saved to a sophisticated XML schema which is created of the PAGE (Page Analysis and Ground truth Elements) Format Framework [12].

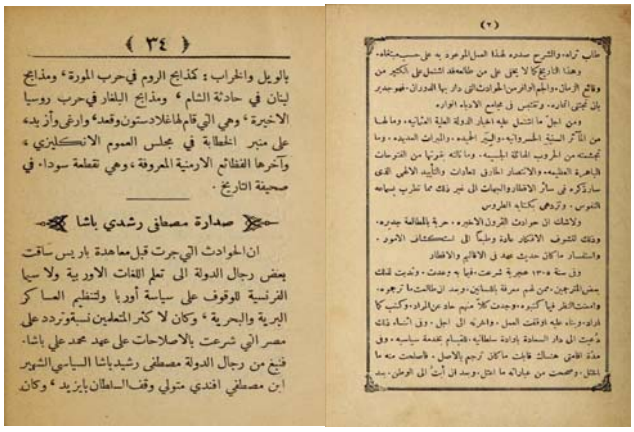
(a) (b)



(c) (d)



(e) (f)



(g) (h)

Fig. 7 Sample book and newspaper pages (a)-(b): Newspapers, (g)-(h): Books

IV. Experiments

We conducted several experiments on the new dataset, using the text line segmentation techniques. First, we ran the algorithms on a sub-set containing images with 0% warping. The results can be seen in Table III and Fig. 8. The second test is done on the same sub-set with a different percentage of warping. The results can be seen in Table IV, Fig. IX, Table V and Fig. 10.

TABLE III
THE SEGMENTATION SUCCESS RATE WITH 0% WARPING

Method	Success Rate
watershed transform	98.5%
Smearing Method	94.9%
Hybrid Approach	93.2%
Projection profile based	90.9%

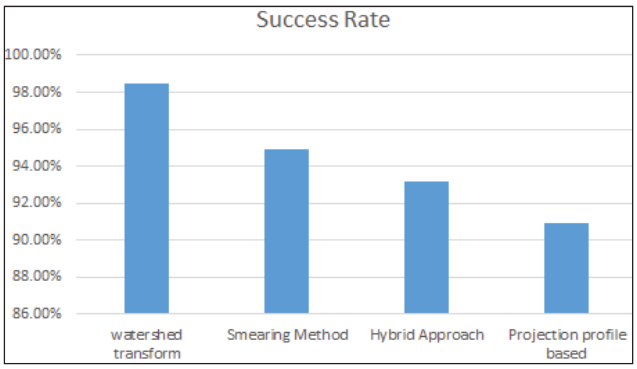


Fig. 8 The success rate with 0% warping

TABLE IV
THE SEGMENTATION SUCCESS RATE WITH25% WARPING

Method	Success Rate
watershed transform	90.50%
Smearing Method	86.90%
Hybrid Approach	81.20%
Projection profile based	75.90%

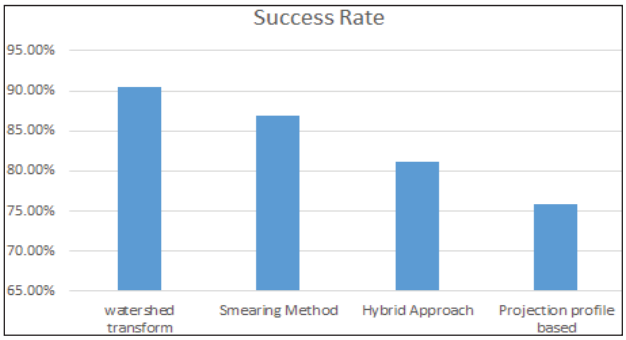


Fig. 9 The success rate with 25% warping

TABLE V
THE SUCCESS RATE WITH 50% WARPING

Method	Success Rate
watershed transform	75.50%
Smearing Method	60.90%
Hybrid Approach	51.20%
Projection profile based	45.90%

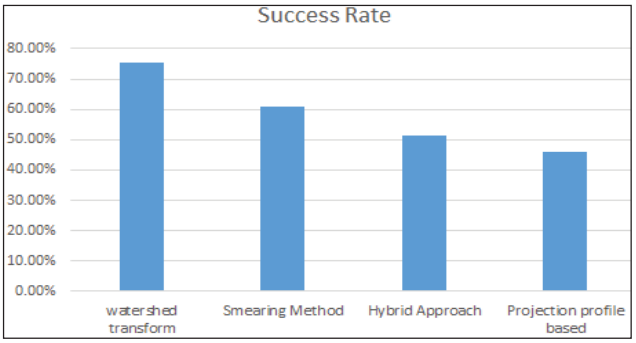


Fig. 10 The success rate with 50% warping

V. Conclusion

We have created a new dataset consisting of 200 pages of archival documents with annotated text line segmentation. A wide span of various times, as well as areas, is included. The dataset includes manuscripts with various degradations and complex layouts. Along with the dataset there is a goal-oriented evaluation scheme based on text line description that has been introduced. In our conclusion, this work presents new challenges as well as a solid basis for completing evaluations for the document layout section. After recognizing some difficulties that occurred, we concentrate our future work to develop to this database and test various Arabic calligraphic styles.

Acknowledgment

The author would like to thank the Pattern Recognition and Image Analysis Research Lab at School of Computing, Science and Engineering University of Salford, UK

References

[1] Yang, P., Antonacopoulos, A., Clausner, C. & Pletschacher, S. Grid-based modelling and correction of arbitrarily warped historical document images for large-scale digitisation. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, 2011. ACM, 106-111.

[2] Lund, W. B. 2014. Ensemble Methods for Historical Machine-Printed Document Recognition
[3] Rahnemoonfar, M. 2010.Correction of arbitrarygeometric artefactsin historical documents. Salford: University of Salford.
[4] M. Pechwitz, S. S. Maddouri, V. M¨argner, N. Ellouze, H. Amiri, et al., "Ifn/enit-database of handwritten arabic words," in Proc. of CIFED, vol. 2pp. 127–136, Citeseer, 2002.
[5] Slimane, F., Ingold, R., Kanoun, S., Alimi, A.M. and Hennebert, J., 2009, July. A new arabic printed text image database and evaluation protocols. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 946-950). IEEE.
[6] Mahmoud, S. A., Ahmad, I., Alshayeb, M., Al-Khatib, W. G., Parvez, M. T., Fink, G. A., Märgner, V. and El Abed, H., 2012, September. KHATT: Arabic offline handwritten text database. In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on (pp. 449-454). IEEE.
[7] Mousa, I. S. 2001. The Arabs in the first communication revolution: development of the Arabic Script. Canadian Journal of communication, 26.
[8] Abuhaiba, I. S. 2003. A discrete Arabic script for better automatic document understanding.
[9] Alromima W, Elgohary R, Moawad IF, Aref M. Applying ontological engineering approach for Arabic Quran corpus: A comprehensive survey. InIntelligent Computing and Information Systems (ICICIS), 2015 IEEE Seventh International Conference on 2015 Dec 12 (pp. 620-627). IEEE.
[10] Suen, C. Y., Nikfal, S., Zhang, B. and Janbi, J., 2017. Characteristics of English, Chinese and Arabic Typefaces. In Advances in Chinese Document and Text Processing (pp. 1-30).
[11] Clausner, C., Pletschacher, S., and Antonacopoulos, A.(2011). Aletheia - an advanced document layout andtext ground-truthing system for production environments.In International Conference on Document Analysis and Recognition. Beijing, China, pp. 48–52.
[12] S. Pletschacher and A. Antonacopoulos, "The PAGE (PageAnalysis and Ground-Truth Elements) Format Framework",Proc. ICPR2008, Istanbul, Turkey, August 23-26, 2010,IEEE-CS Press, pp. 257-260.