An Historical Handwritten Arabic Dataset for Segmentation-Free Word Spotting – HADARA80P

Werner Pantke, Martin Dennhardt, Daniel Fecker, Volker Märgner, Tim Fingscheidt
Institute for Communications Technology
Technische Universität Braunschweig
Braunschweig, Germany
Email: {pantke,fecker,maergner,fingscheidt}@ifn.ing.tu-bs.de
m.dennhardt@tu-bs.de

***Abstract*—In this paper, we present a new and freely available dataset comprising 80 pages of an historical handwritten Arabic document in conjunction with a detailed ground truth for the development and evaluation of segmentation-free word spotting approaches. Besides information on the underlying manuscript and technical details, we introduce a comprehensive list of tags that each word is labeled with. These tags can be used for research on specific issues such as dealing with text in different colors. For comparison of different word spotters, a fixed set of 25 keywords with different properties is included. Furthermore, some specifics of spotting on Arabic manuscripts are discussed. We exemplarily present a state-of-the-art word spotting algorithm in its original and a new extended implementation and evaluate both approaches on the new dataset. For comparison, they are also tested on the widely used George Washington dataset. It is shown that the extended word spotter outperforms the original version in terms of mean average precision on both datasets.**

I. Introduction

Word spotting describes the task of retrieving occurrences of a given keyword in images. In particular for segmentation-free word spotting on scanned documents, it is not only a matter of classifying, i. e., recognizing, pre-segmented word images, but also of locating word occurrences on entire document pages. As there is a trend towards segmentation-free word spotting, there is also the need for unsegmented datasets with detailed information on word level.

There are several widely used handwriting datasets available. The comprehensive IAM handwriting database [1] contains contemporary English sentences written by several writers. The dataset originally targets handwriting recognition, but also provides word segments in form of bounding boxes. The contemporary IfN/ENIT dataset [2] also targets handwriting recognition, but consists of segmented Arabic word images. The historical datasets St. Gall [3] (Latin, 60 pages) and Parzival [4] (medieval German, 47 pages) from the IAM-HistDB collection have already partly been used for word spotting and provide transcriptions on line level. Both datasets do not contain word coordinates on an unsegmented level. One of the most popular datasets for word spotting on historical documents is the George Washington (GW20P) dataset [5] that has been used to evaluate numerous word spotting approaches, among them are [6]–[8]. It provides 20 page images containing historical letters in English and bounding box coordinates of 4856 word occurrences in total.

For word spotting on historical Arabic documents, not much data is available. In [9], an overview of available Arabic datasets is given. None of them contains word coordinates on unsegmented pages, but provides isolated numbers, characters, words, or text lines instead. In [6] and [10], Leydier et al. applied their approaches to two historical Arabic manuscripts, but had to verify the relevance of each retrieved word occurrence manually because of a missing ground truth.

In this paper, we introduce a new dataset, HADARA80P, comprising 80 pages from an historical Arabic manuscript and a detailed ground truth containing polygonal word segments and additional information via tags on word level. The dataset is freely available to the research community. We introduce the new dataset in detail in Section II. In Section III, we briefly describe a state-of-the-art word spotting approach and a new extended version. Subsequently, we apply both spotting systems to the popular GW20P and the new HADARA80P datasets in Section IV and conclude in Section V.

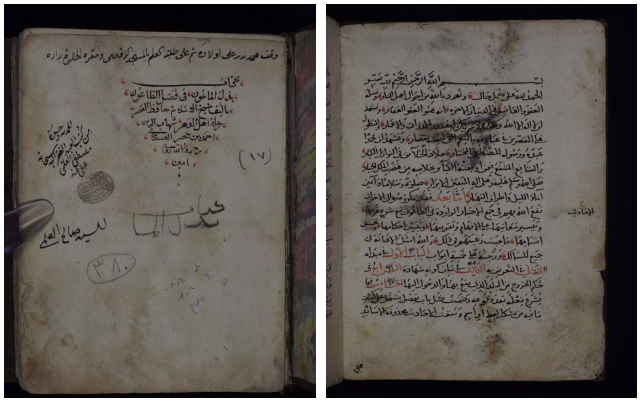
II. The HADARA80P Dataset

In this section, we describe the HADARA80P dataset. It was created within the scope of the HADARA project, which focuses on the analysis of historical handwritten Arabic documents [11]. Therefore, the dataset is based on an historical Arabic manuscript, of which some properties are described in subsection II-A. The data acquisition is outlined in subsection II-B, introducing details about the utilized scanning device as well as about adopted rules of annotation and the utilized annotation tool. Afterwards, the format in which all data are stored are briefly described in subsection II-C, followed by a presentation of a selection of keywords in subsection II-D.

A. The Taaun Book

The HADARA80P dataset is based on the historical handwritten book بَذْلُ المَاعُون فِي فَضْلِ الطَّاعُون *baḏlu ālmāʿūn fī faḍlu āl*ṭ*āʿūn*. The title might be translated to "About the advantage of the pest". The complete book consists of approximately 250 text pages and consists of five chapters. It was published in 06. 833 AH (Islamic calendar), which corresponds to Feb. 1430 AD. The author, Ahmed Ibn Ali Ibn Mohamed Ibn Mohamed Ibn Ali Ibn Mahmoud Ibn Ahmed Ibn Hijr El Kanani El Askalani (short: El Hafid Ibn Hajr El Askalani), lived in Egypt and originated from Palestine. According to side notes in the book itself, the author appears to be also the writer of the book, which may not be the usual case for historical documents. An analysis based on text-independent features for writer identification [12] also reveals that the

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.11



(a) (b)

Fig. 1. First two text pages of the Taaun manuscript – (a) shows information on the author and the book itself, (b) depicts a typical page of this dataset

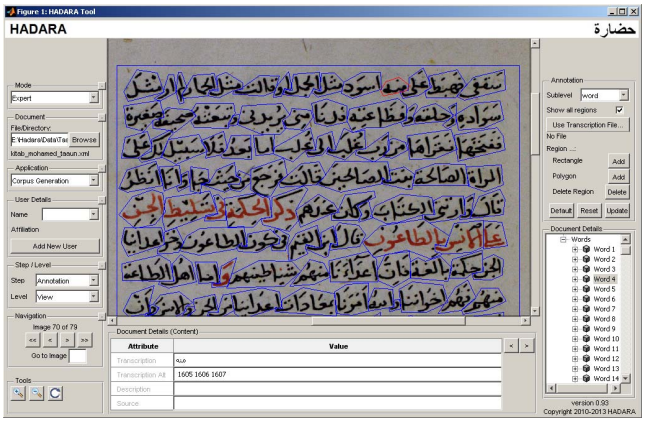


Fig. 2. HADARA annotation tool displaying text block and (polygonal) word segments

whole book, except for the last three text pages, appears to be written by a single writer, which is again not usual in historical documents [13]. The book is about the taaun (pest) disease and its theological meaning, about how to combat this disease and where it appeared in history. In the following, we use *Taaun book* as short name.

The first page of the Taaun book contains a short description, while all the other text pages have one single main text block with relatively few side notes in each case. Fig. 1 shows the first two text pages of the book. Typically for historical Arabic manuscripts, red color in addition to black color is used in the text. Red color is used to emphasize words or text passages, but also to structure the content: Chapter names are mostly written completely in red, while section names are written in red with black diacritics.

B. Data Acquisition

The Taaun book can be found in the private, family-run Al-Budeiri library, which is located in the Old City of Jerusalem. Nowadays, the catalog of this library is online thanks to the MANUMED project and its *Virtual Library of the Mediterranean Sea* [14] funded by the European Union.

For the scanning process, we used the mobile scanning equipment *Traveller's Conservation Copy Stand TCCS 4232* and a *Canon EOS 5D Mark II* fullframe camera with 50 mm macro lens. The lighting conditions were fixed and the resulting scans have a resolution of about 300 dpi containing one document page per image. In order to keep detailed color information of page and text, the camera was configured to output RAW images with a resolution of real 12 bits per color channel. For easier handling, however, we provide standard 48-bit TIFF images with 16 bits per color channel. The cutout of each image has the same dimension of 2882×3650 pixels. All images are stored using the lossless TIFF deflate compression, which leads to a size of about 51 MBytes per image.

The book is annotated and transcribed by native Arabs with the knowledge of Classical Arabic. They used a dedicated tool which interactively supports the annotation process [11]. Fig. 2 shows a screen shot of the graphical user interface of this HADARA annotation tool. A part of a Taaun book page is shown, in which text block and word segments are marked by blue rectangles or polygons, respectively. A single word is highlighted by a red polygon and its transcription is shown in the input fields right under the displayed page image. Besides this annotation mode, also a verification mode is available, with which additional annotators can verify or adapt the annotation.

C. Data Format and Properties

The HADARA80P dataset contains 80 images consisting of one inlay cover and the first 79 text images of the Taaun book. The images are complemented with one XML file containing the ground truth. The annotation section of the ground truth includes segment coordinates for pages, text blocks, and words. All segments are rectangular, except for words, which are described by detailed polygons. Additionally, the annotation includes the UTF-8 transcription of each word. The transcription is narrow, i. e., it resembles exactly what is written, including existent or omitted diacritics.[1] One exception is ligatures, which are described below. This narrow transcription may be used to search for specific writing styles. If such a narrow distinction is not necessary because, e. g., all different styles of a word should be found, the diacritical information can easily be stripped off. In total, the dataset contains 16720 annotated words.[2]

For word segmentation, the annotators were instructed to create word polygons that are relatively close to the characters and contain all corresponding diacritics, but, wherever possible, no strokes from other words. At the same time, they were directed to use as few vertices as possible while meeting the aforementioned requirements. The segmentation of pages meets the following requirements:

1) pages contain all textual information of an image,
2) pages contain as little margin noise[3] as possible w/o conflicting rule 1),
3) pages are as large as possible w/o conflicting rule 2).

The full annotation guidelines are provided with the dataset.

[1] In Arabic, there are optional diacritics such as short vowels and mandatory ones that are needed to distinguish between characters.

[2] Words in side notes, which are mainly written by different writers and occur relatively rarely in this dataset, are not annotated and, therefore, not included in this word count.

[3] For example, the area around a page in which you can see the stacked paper from other pages or the plastic stick that is used to flatten a page (Fig. 1(a)).

For more detailed information about the transcription of a word, we developed a set of tags. Each word is optionally labeled with an arbitrary number of tags listed in Table I. This additional information can be used to select specific words for the training step of a learning-based word spotter or to define which word occurrences from the ground truth are considered to be relevant and, therefore, should be found by a word spotter. So, this information could also be incorporated into the evaluation. Examples of the application of these tags are:

- `redword`, `redpart`: These tags can be used to specifically analyze what happens if the text color of relevant word occurrences does not match the color of the word template.
- `invalid`: A drawback of the narrow transcription is that it can contain words that are invalid according to Arabic language knowledge if the writer made a mistake. In this case, the word is tagged accordingly and can be handled differently.
- `ligature`: Due to UTF-8 limitations, ligatures are not written in the transcription, but the corresponding word is tagged containing the exact character combination. Particular ligatures can be found easily in this dataset.

D. Keyword Set

Besides providing the ground truth, the HADARA80P dataset also includes a selection of 25 cherry-picked keywords that can be used for comparison with other word spotting systems if, e. g., taking each and every word in the dataset as query is computationally not feasible. These 25 keywords consist of words with the highest number of occurrences in the dataset as well as words that may be interesting for the word spotting task due to linguistic properties. For template-based word spotting systems, a template is provided for each keyword, either. All templates are selected following the same rule: The very first keyword occurrence in the document starting from the second text page is picked. The first page is omitted because of word elongations and a generally different writing style as can be seen in Fig. 1(a). Nevertheless, word occurrences on this page are still regarded as relevant and supposed to be found by a word spotting system.

Templates are provided in the following variants: For each keyword, its polygonal coordinates are given that enable a keyword spotter to extract the template image on its own. Additionally, already extracted template images are available. The background of these rectangular template images is filled with the median of all pixel colors present in the keyword polygon, seperately obtained on each color channel. Word spotting systems that require rectangular template images may use these images directly. For both representations, the templates do *not* contain optional diacritics, unlike the word annotation in the ground truth. This enables both learning and template-based approaches to spot for word occurrences in all diacritical variants present in the ground truth. If a spotting system is supposed to distinguish specific writing styles with respect to diacritics, the full keyword polygons are still available in the ground truth.

Table II shows all 25 keywords with their filled bounding box templates and their number of occurrences in the dataset. Due to the fact that the Arabic language enlarges words with

TABLE I. TAGS THAT EACH WORD IN THE GROUND TRUTH IS LABELED WITH WHERE APPROPRIATE

Tag	Example	Description
`invalid`		Word is invalid, i. e., it is not a valid Arabic word due to missing or swapped characters, for instance; the characters can be read visually
`unreadable`	—	Word is not readable (visually) due to noise, for instance, or being crossed out
`stretched`		Word is elongated by a long horizontal line
`shrunk`		Word is shrunk or squeezed to fit into a text line; e. g., at the end of a line, the word often starts above the preceding word from the same text line
`redword`		*Whole* word is written in red color (corresponding diacritics may or may not be red)
`redpart`		Only a *part* of the word (consisting of one or more characters or diacritics) is written in red color
`redfilling`		A character is filled with red color
`numarabic`	—	Word is a number written in Arabic numerals (0, 1, 2, 3, . . .)
`numindic`		Word is a number written in Arabic-Indic numerals (٠,١,٢,٣, . . .)
`punctuation`		Punctuation mark, such as comma, period, slash, question, or exclamation mark
`prosody`		Word is marked with a prosody mark, e.g., a long line (*Maddah*) for prolongated pronunciation
`incomplete`		Word misses diacritic(s) or dot(s)
`enriched`		Word contains additional diacritic(s) or dot(s) that do not belong to the word
`ligature`		Word contains ligature(s), which means that two or more characters are written as a combined character; the mandatory lam-aleph ligature is not tagged here; the single characters of each ligature and not the ligatures themselves are listed next to the tag, grouped by parentheses: example: `ligature` (في) (مح) (لم) for في *fy*, مح *mḥ*, and لم *lm*
`alternative`		Word contains alternative/ancient writing style of at least one character which *cannot* be typed/expressed in the transcription
`individual`		Word contains individual, not commonly used writing style or even spelling mistake that seems to be used consistently throughout the book/manuscript
`abbreviation`	—	Word is an abbreviation
`other`		Other problem/characteristic, e. g., wrapped words

prefixes and suffixes representing, e. g., person, number, tense, and case but also parts of speech such as prepositions and even some conjunctions, it might also be interesting for the word spotting evaluation to allow substring matches. These are matches where the keyword transcription is a substring of other words (ignoring optional diacritics). Of course, not all grammatical variations are possible to be found this way. However, if the keyword is really existent on a page, but just surrounded by other characters, this word occurrence may be regarded as relevant and supposed to be found. For the analysis of this case, there are two pairs of keywords present in the keyword set where one keyword is a substring of the other. KW.24 (طاعون *ṭāʿūn*) is a substring of KW.03 (الطاعون *ālṭāʿūn*) with the latter just having the definite article ال *al*

TABLE II. KEYWORD STATISTICS OF THE HADARA80P DATASET PRESENTING 25 KEYWORDS AND THE CORRESPONDING NUMBER OF RELEVANT OCCURRENCES ACCORDING TO THE GROUND TRUTH (OPTIONAL DIACRITICS ARE IGNORED)

Keyword		# Relevant Occurrences	
Abbreviation	Template Image	Full Match	Substring Match
KW.01		349	370
KW.02		90	131
KW.03		147	154
KW.04		26	30
KW.05		23	42
KW.06		45	46
KW.07		58	115
KW.08		41	44
KW.09		65	73
KW.10		24	24
KW.11		79	177
KW.12		48	50
KW.13		34	48
KW.14		46	46
KW.15		25	25
KW.16		45	54
KW.17		64	68
KW.18		25	38
KW.19		46	47
KW.20		48	64
KW.21		20	20
KW.22		29	44
KW.23		24	24
KW.24		5	169
KW.25		26	26

prefixed. The other keyword pair is an example of a bad word match: KW.02 (سلم *slm*) is a substring of KW.18 (مسلم *mslm*) when no diacritics for short vowels are written, which is often the case.[4] With these optional diacritics, however, it is سَلَم *salam* and مُسلِم *muslim*, respectively, i. e., completely different words with respect to content. In our opinion, the substring match is still valid to identify relevant word occurrences as the written Arabic language itself has this ambiguity. The number of relevant word occurrences in case of allowed substring matches is presented in the last column of Table II. Please note that the following should be kept in mind when allowing substring matches: Due to the fact that Arabic characters change their shape depending on their position inside a word, template-based approaches may have problems finding all of these occurrences, resulting in a lower recall rate.

III. WORD SPOTTING APPROACH

As an application to the HADARA80P dataset, we use an own implementation of the template-based approach presented in [6] and refined in [10], which we further extended. In the following, we refer to this system as the HADARA word spotter. In addition, we test the commercially available word spotting application Ulysse [15], which is the original implementation of [6] and [10]. The method uses gradient angle and magnitude as features, which are compared within automatically identified *zones of interest* (ZOIs) using a *cohesive elastic matching* that is robust against typical variations in handwritten text.

The original word spotting approach can briefly be described as follows [6], [10]: First, all images are converted to grayscale using a pseudo-luminance $L' = L \cdot (1 - S)$ for each pixel, where L is the luminance of the pixel and S its saturation. The word spotting algorithm then finds ZOIs on the template image by locating local maxima of its gradient field curvature. At those positions, ZOI *seeds* are placed and then gradually enlarged as long as the entropy of the gradient angles inside a ZOI rises significantly between two iterations. To reduce the computational costs, the template is only matched against document areas assumed to be interesting, marked by *guides*. The template is propped up against each guide and the template ZOIs are then allowed to be moved slightly in all directions to find a position of minimal distance between the template and the document image inside the ZOIs. The accumulated distance of all ZOIs is assigned to this assumed word occurrence as *score*. Output of the algorithm is an n-best list of areas in the scanned document that resemble the shape of the template image, with n denoting the number of results.

For our own implementation, we modified the ZOI detection to yield better results on high resolution images by not placing ZOI seeds at local maxima of the gradient field curvature, which leads to many ZOIs in insignificant places. Instead, we place seeds where the value of the curvature exceeds a global threshold of 45°. In [6], the allowed displacement of each ZOI on the document image is linked to the average character width and height in the document. The HADARA word spotter allows ZOIs to translate in distances of half their own size in both dimensions instead. Our algorithm is also extended to incorporate two additional steps: Prior to the grayscale conversion, we apply a shading correction to the images that compensates for background variations. For this purpose, we use the bottom-hat transform [16] on the luminance component L^* of a pixel to estimate the local background luminance. The structuring element has to be large enough that it does not fit into textual elements and it depends, therefore, on the mean stroke width of the text. For our images that have a resolution of about 300 dpi, we employ a disc with a radius of 20 pixels as structuring element for the bottom-hat transform. Additionally, after the matching is completed, we check for assumed word occurrences that overlap with others by more than 30 % of their area. Of those pairs, the one possessing the highest score, i. e., greatest distance, is removed. This result filtering tries to avoid returning the same word occurrence in a document several times.

IV. PERFORMANCE EVALUATION

In this section, we test the suitability of the HADARA80P dataset for the word spotting task by comparing both word spotting systems presented in Section III on these data as well as on a standard word spotting dataset. For the evaluation, we follow the complete evaluation protocol of [17], which ensures a stable and fair evaluation process and implicitly punishes

[4] For demonstration purposes, the short vowels are also omitted in the Latin transliteration here.

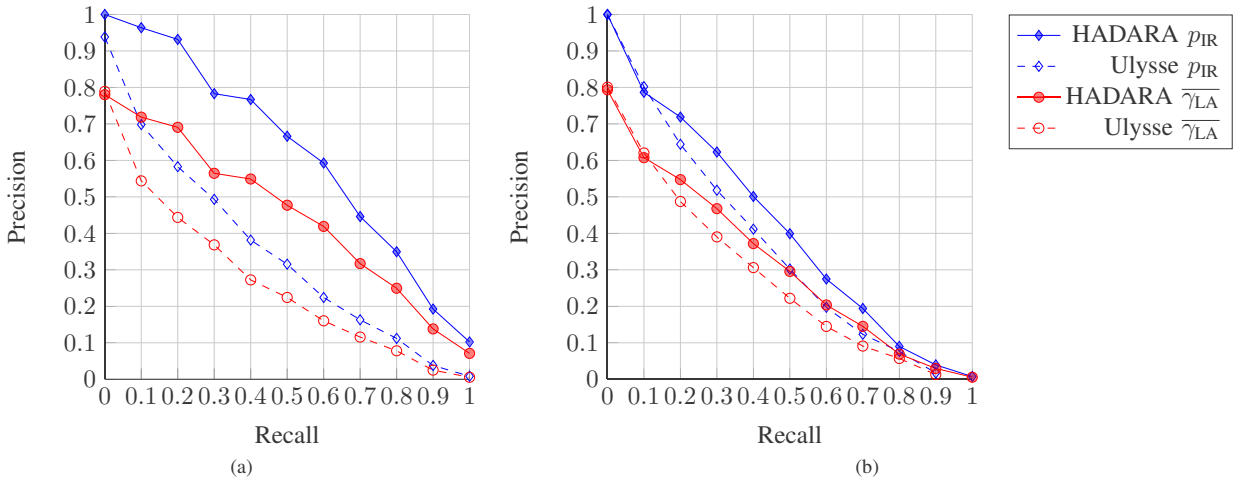


Fig. 3. HADARA and Ulysse word spotting systems applied to (a) GW20P dataset and (b) HADARA80P dataset

cases such as relevant areas of the ground truth being hit multiple times. We also use the precision measures p_{IR} and $\overline{\gamma_{LA}}$ proposed there. The traditional precision measure p_{IR} from the information retrieval domain performs a hard decision whether a retrieved word occurrence is considered relevant or not, i. e., if it is a true or false positive. Here, we require a minimum amount of 1 pixel overlap as relevance threshold. In the second measure $\overline{\gamma_{LA}}$, a true positive does not contribute a binary value to the final precision value, but instead judges the quality of retrieved word locations in a continuous way. So, a match with the retrieved word area being identical to the relevant area from the ground truth results in a perfect location quality of 1. Retrieved parts outside the relevant area and relevant parts that are not covered by the retrieved word area decrease the location quality accordingly. The location quality depends on a parameter c which defines the maximum allowed irrelevant area size as multiplier to the respective relevant word area size. For all of the subsequent experiments, we set $c = 3$. For more details, we refer to [17].

For the evaluation, we utilize the standard 11-point interpolated recall/precision graph [18] and the mean average precision (mAP) [18] value based on the precision measures p_{IR} and $\overline{\gamma_{LA}}$. As defined, both the graph and the mAP value are averaged over all queries.

First, we use the de-facto standard dataset for word spotting in order to show how both word spotting systems presented in Section III perform in general. This is the George Washington dataset (GW20P), which consists of 20 pages containing letters of George Washington [5]. The ground truth provides rectangular word segments. Another important difference to the HADARA80P dataset is the English language and, hence, the Latin alphabet. Both spotting systems were set up to output a ranked list of the best 1000 possible matches according to their internal score. Both systems return rectangular word areas. The same 15 keyword templates as in [6] are used.

As can be seen in Fig. 3(a) and Table III, both systems can be used to spot for keywords. Analyzing the penalty factors of $\overline{\gamma_{LA}}$ reveals why the corresponding curve is lower than using the traditional precision measure: Both systems, but especially the HADARA spotter, return too small word areas compared to the ground truth and are, therefore, punished. Despite their algorithmic relationship, the HADARA word spotter outperforms Ulysse in terms of both employed precision measures.

TABLE III. MEAN AVERAGE PRECISION (MAP) FOR ULYSSE AND HADARA WORD SPOTTING SYSTEMS ON GW20P AND HADARA80P DATASETS USING TWO DIFFERENT PRECISION MEASURES

	GW20P		HADARA80P	
	p_{IR}	$\overline{\gamma_{LA}}$	p_{IR}	$\overline{\gamma_{LA}}$
Ulysse	0.34	0.26	0.35	0.27
HADARA	**0.61**	**0.44**	**0.41**	**0.31**

Second, the same setup as before is used on the new HADARA80P dataset. Both word spotters are applied to unsegmented images containing pages and border noise. The evaluation allows substring matches as described in subsection II-D, considering all word occurrences that are equal to or contain the query word as being relevant and supposed to be retrieved. Obviously, the precision measure $\overline{\gamma_{LA}}$ is affected by allowing substring matches because the relevant areas from the ground truth may be larger than the template. Hence, imprecise matches in terms of both location quality and text quality decrease this precision measure here. The substring matching also implies the possibility that some character shapes are different to those from the respective keyword template as mentioned before, although both tested word spotting systems do not expect other shapes than used in the template. The template images presented in Table II are used, in which optional diacritics are stripped off.[5] The results of this test can be seen in Fig. 3(b) (recall/precision curves) and Table III (mAP values). The HADARA spotter again delivers better results than Ulysse also on this dataset.

In Fig. 4, some example matches of the HADARA word spotter are shown. Fig. 4(a) illustrates a substring match: A word occurrence of KW.03 (الطاعون *ālṭāʿūn*) has been found, although it was searched for the keyword KW.24 (طاعون *ṭāʿūn*). The location quality γ_{LA} for this single match

[5]Actually, the HADARA word spotter only gets keyword template coordinates as input and does the background filling described before itself, whereas Ulysse is fed with (identical) template images.

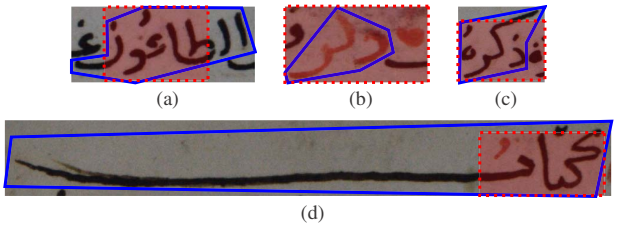


Fig. 4. Example matches using HADARA word spotter – solid polygons illustrate relevant word areas from the ground truth, whereas dotted rectangles depict retrieved word areas from the word spotter

is $0.68 < 1$ and, hence, punished. The keyword template KW.07, which is written in red color, has found red as well as black word occurrences (s. Fig. 4(b)+(c)). However, only 44 of the 115 relevant word occurrences from the ground truth have been found at all, resulting in a recall rate of only 0.38 for this single query. The mean recall rate over all queries is 0.64. Fig. 4(d) shows an elongated word occurrence (tagged with `stretched`) that has correctly been found, but has a low location quality of $\gamma_{\text{LA}} = 0.20 < 1$ because it could not correctly find the word boundaries. The HADARA word spotting system obtained the best results spotting for KW.25 (p_{IR}-mAP: 0.87, recall: 1.00) and the worst results for KW.22 (p_{IR}-mAP: 0.04, recall: 0.27).

V. Conclusion

In this paper, we presented the new HADARA80P dataset for the development and evaluation of word spotting systems targeting historical handwritten Arabic manuscripts. This dataset is freely available to research institutions.[6] We provided details about the underlying manuscript, the data acquisition, and data formats and developed a detailed list of tags on word level that can be used to focus the word spotting development on special issues that arise when spotting on historical Arabic handwritings. The dataset provides unsegmented scanned images of high quality including color information. In addition, segment coordinates of pages, text blocks, and words are provided. Word segments are provided in terms of detailed polygon coordinates. For comparison reasons, we also provide a set of 25 keyword templates with different properties. To the best of our knowledge, this is the first annotated dataset for this purpose and in this level of detail.

Using a state-of-the-art word spotting approach in form of a commercially available original implementation as well as an own implementation with some extensions, we showed that the new dataset can indeed be used to evaluate segmentation-free word spotting approaches. Furthermore, it could be shown on the new dataset as well as on the standard George Washington dataset that our extended word spotting system outperforms the original system. Due to the extensive tag details of this dataset, the research on specific problems of Arabic word spotting may be supported.

Acknowledgment

The authors would like to thank the Al-Budeiri library for the scanning permission and hospitality and Az-Eddine Mostakil for doing a major part of the tedious annotation task as well as for working out the origin of the Taaun book. Additionally, the authors express their gratitude to Abedelkadir Asi for taking part in the scanning process and Fouad Slimane for assisting in finding sensible keywords. A special thank goes to Yann Leydier for providing us with a license of the Ulysse word spotter, adding a command-line interface to it and making it work under Windows and Linux. This work has been funded by the German Research Foundation (DFG) within the scope of the research grant FI 1494/3-2.

[6] http://www.ifn.ing.tu-bs.de/research/data/HADARA80P

References

[1] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *Int. J. Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, 2002.

[2] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri, "IFN/ENIT - database of handwritten Arabic words," in *Proc. Colloque international francophone sur l'écrit et le document (CIFED 2002)*, Hammamet, Tunisia, 2002, pp. 129–136.

[3] A. Fischer, V. Frinken, A. Fornes, and H. Bunke, "Transcription alignment of Latin manuscripts using hidden Markov models," in *Proc. Workshop on Historical Document Imaging and Processing (HIP 2011)*. Beijing, China: ACM, 2011, pp. 29–36.

[4] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934–942, May 2012.

[5] V. Lavrenko, T. M. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents," in *Proc. Int. Workshop on Document Image Analysis for Libraries (DIAL 2004)*, Palo Alto, CA, USA, 2004, pp. 278 – 287.

[6] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Text search for medieval manuscript images," *Pattern Recognition*, vol. 40, no. 12, pp. 3552–3567, 2007.

[7] L. Rothacker, M. Rusinol, and G. A. Fink, "Bag-of-features HMMs for segmentation-free word spotting in handwritten documents," in *Proc. Int. Conf. Document Analysis and Recognition (ICDAR 2013)*, Washington DC, USA, 2013, pp. 1305–1309.

[8] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 211–224, 2012.

[9] S. N. Srihari and G. R. Ball, "An assessment of Arabic handwriting recognition technology," in *Guide to OCR for Arabic Scripts*, V. Märgner and H. El Abed, Eds. Springer-Verlag London, 2012, ch. 1, pp. 3–34.

[10] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition*, vol. 42, no. 9, pp. 2089–2105, 2009.

[11] W. Pantke, V. Märgner, D. Fecker, T. Fingscheidt, A. Asi, O. Biller, J. El-Sana, R. Saabni, and M. Yehia, "HADARA – A software system for semi-automatic processing of historical handwritten Arabic documents," in *Proc. Archiving Conf. 2013*, Washington DC, USA, April 2013, pp. 161–166.

[12] D. Fecker, A. Asi, V. Märgner, J. El-Sana, and T. Fingscheidt, "Writer identification for historical Arabic documents," in *Proc. Int. Conf. Pattern Recognition (submitted)*, Stockholm, Sweden, August 2014.

[13] R. Altman, "The illusion of one writer in historical documents and its effect on automating writer identification," in *Proc. Conf. of Int. Graphonomics Society*, Dijon, France, September 2009.

[14] EU MANUMED project. Virtual library of the Mediterranean Sea. [Online]. Available: http://data.manumed.org/

[15] CoReNum. Ulysse – universal search engine. [Online]. Available: http://www.corenum.com/products/ulysse/

[16] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*, 3rd ed. Prentice Hall, Aug. 2007.

[17] W. Pantke, V. Märgner, and T. Fingscheidt, "On evaluation of segmentation-free word spotting approaches without hard decisions," in *Proc. Int. Conf. Document Analysis and Recognition (ICDAR 2013)*, Washington DC, USA, 2013, pp. 1300–1304.

[18] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008, vol. 1.