Published version: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6628700

Novel Sub-Character HMM Models for Arabic Text Recognition

Irfan Ahmad
Information and Computer Science
King Fahd University of Petroleum & Minerals (KFUPM)
Dhahran, Saudi Arabia
irfanics@kfupm.edu.sa

Leonard Rothacker, Gernot A. Fink
Faculty of Computer Science
Technische Universität Dortmund
Dortmund, Germany
{leonard.rothacker, gernot.fink}@tu-dortmund.de

Sabri A. Mahmoud
Information and Computer Science
King Fahd University of Petroleum & Minerals (KFUPM)
Dhahran, Saudi Arabia
smasaad@kfupm.edu.sa

***Abstract*—Hidden Markov Model (HMM) is one of the most widely used classifier for text recognition. In this paper we are presenting novel sub-character HMM models for Arabic text recognition. Modeling at sub-character level allows sharing of common patterns between different contextual forms of Arabic characters as well as between different characters. The number of HMMs gets reduced considerably while still capturing the variations in shape patterns. This results in a compact and efficient recognizer with reduced model set and is expected to be more robust to the imbalance in data distribution. Experimental results using the sub-character model based recognition of handwritten Arabic text as well printed Arabic text are reported.**

Keywords— Arabic text recognition; Hidden Markov Models; Sub-character HMMs; Parameter sharing; OCR

I. INTRODUCTION

Text recognition is an active area of pattern recognition research. Researchers have tried various approaches for text recognition in terms of preprocessing, feature extraction, use of classifiers and post processing. Segmentation verses segmentation free recognition; Structural as well as statistical features; Rule based classifiers, HMMs, Neural networks, SVMs, and hybrid models, have been used for recognition purposes. HMMs are one of the most successful and widely used classifier in the area of text recognition [1]. One of the major benefits of using HMMs is that explicit segmentation of the text into recognition units (like characters or strokes) is not required. Most of the other classifiers do require some sort of segmentation of text.

The general trend in HMM based techniques is to use a sliding window [1]. This technique is used to convert the 2–D word image into 1–D sequence of features required by HMM. Different types of features are extracted from the sliding windows. The sliding window is normally overlapped with the previous window. Different overlapping was used by researchers [2]. The sliding window may be divided into vertical overlapping/non–overlapping cells. The width of the sliding window ranged from one pixel to 16 pixels [3]. The features are estimated for the window as a whole and/or the cells within the sliding window.

Arabic is one of the Semitic languages and is ranked as the fourth most widely spoken language in the world [4]. It is spoken by more than 220 million people and is the official language of 22 countries. Arabic text recognition is relatively new as compared to other major scripts like Latin and Chinese. Recently there has been a surge in research in Arabic text recognition thanks to many interesting and challenging competitions held in the area of Arabic text recognition.

As with the case of other language scripts, different classifiers have been used for Arabic text recognition purpose as well. Among the most widely used are HMMs [5–8], Neural Networks [9], [10], SVMs [11], [12], and rule based classifiers [13]. Most of the classifiers apart from HMM were used mainly in holistic recognition tasks like isolated character or digit recognition and if used for word or unconstrained text recognition, some level of segmentation was required. Arabic script is inherently cursive (both in printed and handwritten form) and as such the issues with segmentation are ever more concerning in the case of Arabic text recognition. A comprehensive survey on Arabic handwritten text recognition can be found in [14]. A more recent survey was presented in [15].

Most researchers using HMM for text recognition adapted their existing recognizer for the Arabic script (e.g. [16]). Among the important adaptations include setting up a right-to-left HMM instead of left-to-right HMMs (used for Latin scripts) and selection of basic HMMs units. Using characters HMMs was the natural choice for most researchers.

Although a lot of research has been done in HMM based Arabic text recognition, we believe that there is still room for exploring the peculiarities of Arabic script and using those observations to better adapt HMM for the task. Among the possible areas of adapting HMMs for Arabic text recognition is the selection of the basic HMM units.

There are different characters in Arabic language which share common base shape but differ based on number and position of dots. Moreover a character in Arabic script can take different forms based on its position in the word. Parts of these character shapes are similar for the different forms of the same character and also between the different characters and their contextual forms. The present paper reports the investigation of this property of the Arabic script for better modeling in HMM by proposing a novel sub-character based HMM models.

The rest of the paper is organized as follows: In Section II we present the related work on using sub-character and sub-stroke HMMs for text recognition. In Section III we present details on the peculiarities of the Arabic script and discussions on how these peculiarities can be exploited for HMM based text recognition. In Section IV we present the details of our sub-character HMM modeling for Arabic text recognition. In Section V we present the experiment results and discussions on

sub-character HMM based text recognition. Finally in Section VI, we present the conclusion of our work.

II. Related work on sub-character and sub-stroke HMM model based text recognition

Some interesting works on sub-character and sub-stroke HMMs based recognition are available in literature for online text recognition (mainly for the East-Asian scripts like Kanji). Nakai et al. presented sub-stroke HMMs based recognition system for Kanji characters [17]. Several motivations were stated for using sub-stroke HMMs as opposed to whole character HMMs including; compact system with less number of models and dictionary size, faster recognition due to efficient sub-stroke network search, and less training data requirement. A set of 25 sub-strokes were identified and modeled. It was stated that the presented 25 sub-strokes can represent different Kanji characters using a dictionary defining the character structures. A hierarchical dictionary was defined where the elementary units are the sub-strokes defining the strokes which in turn defines the sub-Kanji characters finally defining the Kanji characters. Automatic generation of this dictionary was presented in [18]. The recognition results showed no improvement in terms of recognition results (actually the results for sub-stroke HMMs were a little lower) but the benefits of sub-stroke HMMs (low memory requirement, efficient decoding etc.) might outweigh in some situations.

Hu et al. presented sub-character HMM models for online handwriting recognition of isolated digits, characters and isolated words [19], [20]. A character (or digit) was constructed by concatenating sub-character models based on a character lexicon. The main motivation stated for using sub-character models instead of character models was the reduction of model set and less training data requirement due to fewer basic HMMs. However they mentioned that although sub-character HMM model based recognizer will be efficient, the recognition results will not necessarily be higher and in some cases (where large amount of training data is available) may end up lower.

Tokuno et al. presented an interesting work on sub-stroke HMM based online recognition of cursive Kanji and Hiragana characters [21]. They argued that there are more than 6000 characters in Kanji and Hiragana scripts and modeling each character would not be practical because of huge memory requirement and insufficient training data for each unique character. They mentioned that any Kanji character can be modeled by concatenating 25 sub-stroke HMMs. This leads to efficient recognizer in terms of memory and computation time. Further they experimented with context dependent sub-stroke models to capture the variation of sub-strokes due to its context (its adjacent sub-strokes). They employed Successive State Splitting (SSS) algorithm to reduce the number of models by sharing similar states of the context dependent sub-strokes. This lead to increase in recognition accuracy when compared to 'whole character HMM models' as well as context independent sub-stroke HMMS.

The sub-stroke and sub-character HMM based recognition presented above are related to online text recognition. In the current work, we explore the sub-character HMM for offline text recognition of Arabic script. According to the best knowledge of the authors, the sub-character HMM based text recognition (without explicit segmentation) was never proposed for the task of offline text recognition. The concept of sub-character/strokes in offline recognition task is different than online recognition as the temporal information is not available in the case of offline text recognition which makes the task somewhat more difficult. Moreover our reported techniques can be easily adapted to apply on online Arabic text recognition task which might prove to be more promising.

III. Peculiarities of The Arabic script

The Arabic script has 28 basic characters. In addition to these basic characters, there are additional derivatives to some of these characters due to the presence of special diacritics like *Hamza* and *Maddah*. Some character combinations lead to special shapes (ligatures) which visually look much different than the combination of constituent characters.

A character in Arabic script can usually take four different forms depending on its position in the word (by position we mean a character appearing in text ***alone***-without any joining character; ***beginning***-where a character is the first character of a word, ***middle***-where a character is preceded by a character in a word and at least one character follows it in the word, and ***ending***-where a character is the last character in a word). Fig. 1 illustrates the different character shapes for a sample character (*Seen* <س>).This property of the script leads to issues when using characters as basic HMM units. There is huge variability within character shapes which seems difficult for HMM to generalize.

Characters (Latin)	Position	Shape	Example
س (*Seen*)	Alone	س	ناس
	Beginning	سـ	سلام
	Middle	ـسـ	انسان
	Ending	ـس	مجلس

Fig. 1. Different character shapes for a sample Arabic character along with illustrative examples.

Six of the 28 characters (like character *Alif* <ا>, *Waw* <و>, and *Dal*<د>) do not take four different forms but instead take only two forms (i.e. *alone* and *ending*). In such cases even if the character has an actual *middle* or the *beginning* position in a word, no character can join it and as a result, instead of taking the *beginning* position it will take the *alone* position and *ending* position instead of *middle* position. Fig. 2 provide illustrative examples using a sample Arabic character (*Dal*).

Characters (Latin)	Position	Shape	Example
د (*Dal*)	Alone	د	اسود
	Beginning	Takes alone shape	دنيا
	Middle	Takes ending shape	مدرسة
	Ending	ـد	اسد

Fig. 2. An illustrative example of a character in Arabic taking only two contextual shapes.

The variability in character shapes depending on the character and its position led some researchers to use individual character shapes as HMM unit instead of characters as HMM units. Moreover a popular approach is to merge the different character shapes to the representing character post-recognition as they essentially represent the same character. This modification in HMM setup generally resulted in improvements in recognition results. Although this lead to better recognition in most cases, it leads to another problem. The number of basic HMMs grew almost four-fold. Having more models means the need for more data to estimate the parameters and limited data is a common problem. Moreover there are some character shapes whose occurrence in Arabic text is very rare and as such only few samples are available in the training data to train their models. To find a compromise between character HMMs and character-shape HMMs we are proposing sub-character HMMs which can uniquely model all the shape variations and the total HMM models are still comparable to character HMM based recognition.

IV. Novel Sub-Character HMM Models For Arabic Text Recognition

The core idea is to have character segments as HMMs instead of character or character-shapes as HMMs. The justification for this approach is the observation that there are similar patterns between different forms of a character and as well as between different characters in Arabic script. Fig. 3 illustrates this observation using sample characters and their shapes.

Characters	Alone	Beginning	Middle	Ending
س	س	سـ	ـسـ	ـس
ش	ش	شـ	ـشـ	ـش
ص	ص	صـ	ـصـ	ـص
ض	ض	ضـ	ـضـ	ـض

Fig. 3. Some Arabic characters along with their different contexual shapes demonstrating the the presence of similar patterns.

Many of the similar and dissimilar patterns can be captured in horizontal direction. This is important as sequencing the 2-D text image is done by moving the sliding window horizontally. As a result of this sub-character splitting, we achieve parameter sharing as similar patterns among characters and character forms is represented by same HMM and the patterns which are unique to a character or character form is represented by different HMMs. This leads to significant reduction in the number of HMMs while at the same time capturing the visual variability. Fig. 4 shows four common Arabic characters along with its different shapes. By using character-shape HMMs, one needs16 different HMMs for these four characters whereas by using our sub-character HMM models, we reduce the number of HMMs to five (plus the special connector and space model shared by other characters as well).

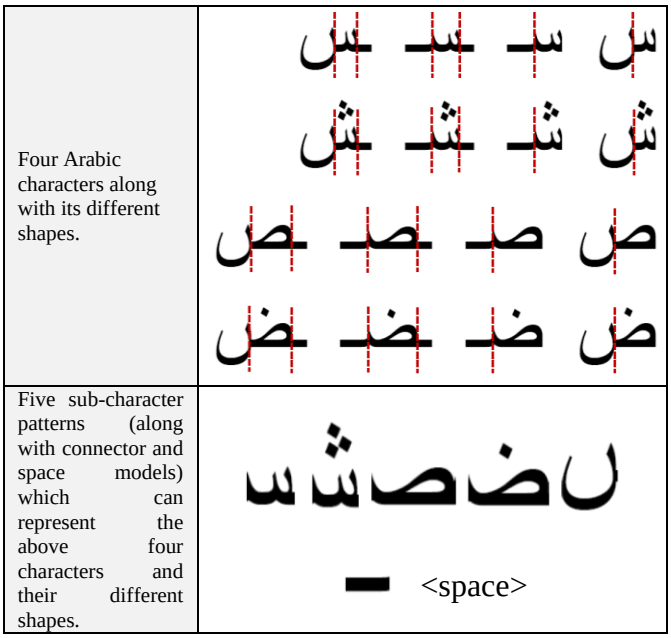

Fig. 4. An example of sub-character modeling reducing the number of basic HMM models.

Once we list all the unique patterns (at sub-character level) in Arabic script using the domain knowledge of the Arabic script, creating the complete HMM structure is relatively straight forward by simply concatenation the resultant sub-character models to form characters and its various contextual forms. Fig. 5 illustrates one possible HMM structure for Arabic text recognition at character level for an example character (*Seen*). More levels in the dictionary structure can be added in the case of lexicon based recognition or other specific recognition tasks.

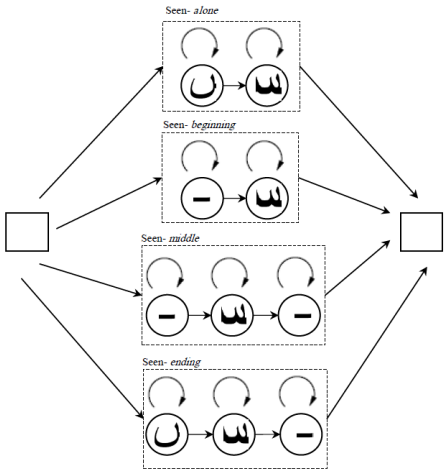

Fig. 5. HMM structure for Arabic character (*Seen*) using the proposed sub-character model.

It is important to mention here that the proposed sub-character HMM models does not need any explicit segmentation of characters into sub-characters as with the case of other segmentation based approaches as discussed by Lorigo et al. in [14].

V. Experiments And Discussion

To study the usefulness of sub-character based HMM modeling for Arabic text recognition; we carried out

experiments for two tasks; handwritten Arabic text recognition, and printed Arabic text recognition. Below we present the details on the conducted experiments and the results.

A. *Experiments related to handwritten Arabic text recognition*

For the task of handwritten Arabic text recognition, we evaluated our sub-character HMM based recognizer on IFN/ENIT database [22] of 32,492 handwritten word images of Tunisian cities and towns that are divided into subsets a – e. In order to compare the results of sub-character HMMs with character-shape HMMs based recognition, we use Bag-of-Features HMMs [23] as our baseline system. The basic idea behind Bag-of-Features HMMs is to integrate Bag-of-Features representations, know from numerous applications in the computer vision domain [24] with HMMs for segmentation-free handwriting recognition. These feature representations have the advantage that they are estimated from sample data. Therefore, they are automatically adaptable to the problem domain. The complete details on the recognizer including the pre-processing, feature extraction, training and evaluation process are presented in [23]. The recognition results of this baseline system is comparable to the current state-of-the-art (cf. Table V, [25]). The only difference between the two modeling approaches are that the baseline system uses 178 character-shape models as elementary units while in the proposed model these are built on the basis of 108 sub-character models. Table I summarizes the results of the experiments on following *training – test* configurations: *abc-d*, and *abcd-e*. The recognition results are shown in terms of word error rate (WER). It can be seen from the table that the recognition results for sub-character HMMs system are almost the same (a slightly higher WER) as the character-shape HMMs system. Nevertheless the sub-character system has much fewer HMMs and the total number of states. This in itself is a big gain resulting in a compact and efficient recognizer. The results also support the findings of other researchers that sub-character HMMs based recognition may not increase the recognition accuracy but does lead to significantly fewer models resulting in a compact system [17] [19], [20].

TABLE I. RESULTS (WER) FOR HANDWRITTEN TEXT RECOGNITION

System	No of HMMs	WER *(Statistical significance)*	
		Exp. Config. abc-d	***Exp. Config. abcd-e***
Character-Shape HMM	178	3.50 *(±0.4)*	7.58 *(±0.7)*
Sub-Character HMM	108	3.62 *(±0.4)*	8.32 *(±0.7)*

B. *Experiments related to printed Arabic text recognition*

For the task of printed Arabic text recognition, we used a database consisting of 6472 training text-line images and 1414 test text-line images. The database is the printed text version of the KHATT database [26] scanned at 300 DPI using a single font. For the experiments, we used the ESMERALDA framework of semi-continuous HMMs for the task of character recognition with 18-dimensional statistical features (pixel density features from the image and its horizontal and vertical edge derivatives). We experimented with two systems. The first system is a character-shape HMM system and the second system in the sub-character HMM system. The summary of the experimental results are presented in Table II. The recognition results are shown in terms of character error rate (CER). We can see from the table that character-shape HMM system and sub-character HMM system have almost the same recognition results with sub-character HMM system showing slightly improved results which are statistically significant at 95% confidence level. More important than the slightly better recognition results is the fact that the sub-character HMM system has significantly less number of HMMs as compared to character-shape HMM system. These results confirm the advantage of sub-character HMM models for text recognition task.

TABLE II. RESULTS FOR PRINTED TEXT RECOGNITION

System	No of HMMs	CER	Statistical Significance
Character-Shape HMM	151	6.51	±0.2
Sub-Character HMM	93	5.98	±0.2

Based on the experiments conducted on two different text recognition tasks we conclude that sub-character HMMs are an interesting possibility for offline text recognition for scripts where similar patterns between characters and characters forms can be explored. The sub-character based HMM recognition system is compact and efficient and most likely will behave robustly in situations where training data is inadequate or the character distribution is imbalanced. Based on our experimental results, the sub-character HMM based system showed slightly better recognition results as compared to character-shape HMM system on printed text recognition task. This may be attributed to the fact that patterns are more consistent in printed text whereas in handwritten text, the variability is much higher. Thus testing contextual sub-character HMMs (something similar to the work in [21]) seems an interesting future work. Moreover it will also be interesting to investigate this approach for online Arabic text recognition. It can be noted from Table 1 and Table 2 that the number of unique sub-character HMMs (as well as character-shape HMMs) is different in handwritten text recognition task as compared to the printed text recognition task. This is due to the presence of many ligatures (like *Lam-Alif*, *Lam-Ha-Meem*) in handwritten text database i.e. IFN/ENIT (each ligature is modeled as separate HMM). Moreover characters with special diacritic (*Shadda*) were also modeled as separate HMMs in the case of handwritten task due to the presence of large number of them in the IFN/ENIT database.

VI. CONCLUSION

Text recognition is an interesting and open research area in the field of pattern recognition. HMMs are widely used classifier for text recognition. A lot of research has already been carried out in advancing the use of HMM in this area of research. In this paper we presented the sub-character HMM based text recognition where the basic HMM units are sub-characters instead of characters or character-shapes. Sub-character HMMs exploits the similar patterns occurring in different characters or different contextual forms of a character which is a prominent characteristic of the Arabic script. Characters can be constructed by simply concatenating the sub-character HMMs according to a predefined dictionary. This allows for parameter sharing between characters and results in a compact recognizer having much less models and thus is more efficient. Two set of experiments were conducted to demonstrate the usefulness of sub-character HMMs in Arabic

text recognition task. More experiments need to be carried out to assess the possible robustness of the resultant system. Although the recognition results are not much different than other common HMM setups nevertheless the benefits of using sub-character HMMs are huge in terms of compact and efficient recognizer (with much fewer HMMs). Investigating contextual sub-character HMMs seems an interesting future work.

ACKNOWLEDGMENT

The authors would like to acknowledge King Fahd University of Petroleum and Minerals (KFUPM) for supporting this research.

REFERENCES

[1] T. Plötz and G. A. Fink, "Markov models for offline handwriting recognition: a survey," *International journal on document analysis and recognition*, vol. 12, no. 4, pp. 269–298, 2009.

[2] Y. Kessentini, T. Paquet, and A. M. Ben Hamadou, "Off-line handwritten word recognition using multi-stream hidden Markov models," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 60–70, 2010.

[3] S. Mozaffari, K. Faez, V. Märgner, and H. El-Abed, "Lexicon reduction using dots for off-line Farsi/Arabic handwritten word recognition," *Pattern Recognition Letters*, vol. 29, no. 6, pp. 724–734, 2008.

[4] "Ethnologue , Languages of the world." [Online]. Available: http://www.ethnologue.com/.

[5] M. Pechwitz and V. Maergner, "HMM based approach for handwritten Arabic word recognition using the IFN/ENIT-database," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, 2003, p. 890.

[6] A. Benouareth, A. Ennaji, and M. Sellami, "Semi-continuous HMMs with explicit state duration for unconstrained Arabic word modeling and recognition," *Pattern Recognition Letters*, vol. 29, no. 12, pp. 1742–1752, 2008.

[7] R. Al-Hajj Mohamad, L. Likforman-Sulem, and C. Mokbel, "Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1165–1177, 2009.

[8] S. Saleem, H. Cao, K. Subramanian, M. Kamali, R. Prasad, and P. Natarajan, "Improvements in BBN's HMM-based offline Arabic handwriting recognition system," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, 2009, pp. 773–777.

[9] A. Amin, H. Al-Sadoun, and S. Fischer, "Hand-printed Arabic character recognition system using an artificial network," *Pattern recognition*, vol. 29, no. 4, pp. 663–675, 1996.

[10] N. Farah, M. T. Khadir, and M. Sellami, "Artificial neural network fusion: Application to Arabic words recognition," in *ESANN 2005 Proc., Eur. Symp. Artificial Neural Networks*, 2005, pp. 27–29.

[11] J. Sadri, C. Y. Suen, and T. D. Bui, "Application of Support Vector Machines for recognition of handwritten Arabic/Persian digits," in *Second Conference on Machine Vision and Image Processing & Applications (MVIP 2003)*, 2003, pp. 300–307.

[12] S. A. Mahmoud and S. M. Awaida, "Recognition of off-line Handwritten Arabic (Indian) Numerals using Multi-Scale Features and Support Vector Machines vs. Hidden Markov Models," *The Arabian Journal for Science and Engineering*, vol. 34, no. 2B, pp. 429–444, 2009.

[13] M. T. Parvez and S. A. Mahmoud, "Arabic handwriting recognition using structural and syntactic pattern attributes," *Pattern Recognition*, vol. 46, no. 1, pp. 141–154, 2013.

[14] L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 5, pp. 712–24, May 2006.

[15] M. T. Parvez and S. A. Mahmoud, "Offline arabic handwritten text recognition: A Survey," *ACM Comput. Surv.*, vol. 45, no. 2, pp. 23:1–23:35, Mar. 2013.

[16] M. P. Schambach, J. Rottland, and T. Alary, "How to convert a Latin handwriting recognition system to Arabic," *Proceedings of the ICFHR*, pp. 265–270, 2008.

[17] M. Nakai, N. Akira, H. Shimodaira, and S. Sagayama, "Substroke approach to HMM-based on-line Kanji handwriting recognition," in *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, 2001, pp. 491–495.

[18] M. Nakai, H. Shimodaira, and S. Sagayama, "Generation of hierarchical dictionary for stroke-order free Kanji handwriting recognition based on substroke HMM," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, 2003, pp. 514 – 518 vol.1.

[19] J. Hu, S. Gek Lim, and M. K. Brown, "Writer independent on-line handwriting recognition using an HMM approach," *Pattern Recognition*, vol. 33, no. 1, pp. 133–147, 2000.

[20] J. Hu, S. G. Lim, M. K. Brown, and others, "HMM based writer independent on-line handwritten character and word recognition," in *Proc. of IWFHR*, 1998, vol. 6, pp. 143–155.

[21] J. Tokuno, N. Inami, S. Matsuda, M. Nakai, H. Shimodaira, and S. Sagayama, "Context-dependent substroke model for HMM-based on-line handwriting recognition," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, 2002, pp. 78–83.

[22] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri, "IFN/ENIT - Database of Handwritten Arabic Words," in *7th Colloque International Francophone sur l'Ecrit et le Document , CIFED 2002*, 2002, pp. 129–136.

[23] L. Rothacker, S. Vajda, and G. A. Fink, "Bag-of-Features Representations for Offline Handwriting Recognition Applied to Arabic Script," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, 2012, pp. 149–154.

[24] S. O'Hara and B. A. Draper, "Introduction to the Bag of Features Paradigm for Image Classification and Retrieval," *Computing Research Repository*, vol. arXiv:1101, 2011.

[25] V. Margner and H. E. Abed, "ICDAR 2011 - Arabic Handwriting Recognition Competition," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 1444–1448.

[26] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Margner, and H. EL Abed, "KHATT: Arabic Offline Handwritten Text Database," in *2012 International Conference on Frontiers in Handwriting Recognition (ICFHR-2012)*, 2012, pp. 447–452.