Journal of
Imaging

MDPI

Article

A Holistic Technique for an Arabic OCR System

Farhan M. A. Nashwan [1], Mohsen A. A. Rashwan [1], Hassanin M. Al-Barhamtoshy [2], Sherif M. Abdou [3,*] and Abdullah M. Moussa [1]

[1] Department of Electronics and Electrical Communications, Cairo University, Giza 12613, Egypt; far_nash@hotmail.com (F.M.A.N.); mrashwan@rdi-eg.com (M.A.A.R.); a.m.moussa@ieee.org (A.M.M.)
[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; hassanin@kau.edu.sa
[3] Faculty of Computers & Information, Cairo University, Giza 12613, Egypt
* Correspondence: s.abdou@fci-cu.edu.eg; Tel.: +20-10-2661-4479

Received: 30 October 2017; Accepted: 22 December 2017; Published: 27 December 2017

Abstract: Analytical based approaches in Optical Character Recognition (OCR) systems can endure a significant amount of segmentation errors, especially when dealing with cursive languages such as the Arabic language with frequent overlapping between characters. Holistic based approaches that consider whole words as single units were introduced as an effective approach to avoid such segmentation errors. Still the main challenge for these approaches is their computation complexity, especially when dealing with large vocabulary applications. In this paper, we introduce a computationally efficient, holistic Arabic OCR system. A lexicon reduction approach based on clustering similar shaped words is used to reduce recognition time. Using global word level Discrete Cosine Transform (DCT) based features in combination with local block based features, our proposed approach managed to generalize for new font sizes that were not included in the training data. Evaluation results for the approach using different test sets from modern and historical Arabic books are promising compared with state of art Arabic OCR systems.

Keywords: Arabic OCR systems; holistic OCR approach; holistic OCR features; lexicon reduction

1. Introduction

Cursive scripts recognition has traditionally been handled by two major paradigms: a segmentation-based analytical approach and a word-based holistic approach. In the analytical approach, the input word is treated as a sequence of units (usually characters). Each unit is then individually recognized [1–4]. This approach has several disadvantages. The segmentation of cursive words is a challenging task and any errors in that process will increase the errors in the following recognition step. Also, many of the used fonts for cursive scripts extensively use ligatures where two or more letters are joined as a single glyph, which complicates the character level segmentation. Figure 1 shows some challenging samples of Arabic words.

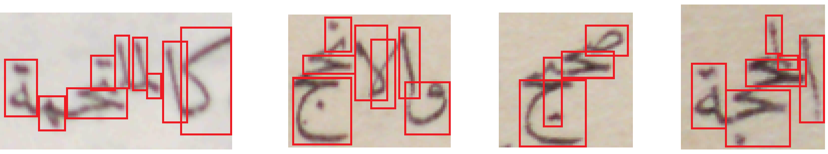

Figure 1. Some examples of Arabic words that contain ligatures with manually segmented characters.

Cursively written word cannot be recognized without being segmented and cannot be segmented without being recognized [5]. This phenomenon, known as Sayre's paradox, pushes the community to search for more effective solutions to tackle the problem of classification. A more direct and efficient methodology can be provided using holistic recognition [6]. Holistic approach handles the whole word as a unified unit. A global feature vector is calculated for the indivisible input word sample which is then utilized to classify the word against a stored lexicon of words. Holistic recognition is inspired from what is known as the word superiority effect, which states that people have better recognition of letters presented within words as compared to isolated letters and to letters presented within non-words [7]. Holistic paradigms are not only effective, but also have the ability to maintain certain effects which are special to the class under operation such as coarticulation effects [8].

Several previous research efforts have investigated the holistic approach for Arabic cursive script recognition for both printed and handwritten types. Erlandson et al. [9] reported a word-level recognition system for machine-printed Arabic. They used an image-morphological based vector of features such as dots and hamzas, the direction of segments, the junctions and endpoints, direction of cavities, holes, descenders and intra-word gaps. All these features are computed for a query word image in the recognition phase and are matched against a pre-computed database of vectors from an Arabic words lexicon and that system achieved a word recognition rate of 65%. This accuracy was achieved with the integration of a lexicon pruning subsystem that is based on another recognition method that was developed under the same project for a training set of 8436 word images scanned at 300 dpi.

Al-Badr et al. [10] developed an Arabic holistic word recognition system based on a set of shape primitives that are detected with mathematical morphology operations. That system was trained using a single font with three types of documents: ideal (noise-free), synthetically degraded and scanned. The used feature extraction operators were very sensitive to the scanning noise and the degraded low resolution documents. That system achieved a recognition rate of 99.4% for noise-free documents. For synthetically degraded documents, the system accuracy decreased to 95.6% and to 73% for scanned documents. All these evaluations were performed using a limited lexicon that contained 4317 words [10].

Khorsheed and Clocksin [11] presented a technique for recognizing Arabic cursive words from scanned images of text by transforming each word in a certain lexicon into a normalized polar image, and then applied a two-dimensional Fourier transform to that polar image. Each word is represented by a template that includes a set of Fourier's coefficients, and for recognition, the system used a normalized Euclidean distance that measures the distance between the word under test and those templates. That system achieved a recognition rate of 90% for a lexicon size of 145 words and used 1700 word samples for training.

To get better performance, Khorsheed [12] presented a new system based on Hidden Markov Models (HMMs). In that system, each word was represented by a single HMM. The word models were trained using the word sample Fourier's spectrum. The experiments were conducted on four fonts, and the reported results are for Simplified Arabic and Arabic Traditional fonts only. The system achieved a higher recognition rate compared to the template-based recognizer. The highest achieved results for both fonts are: 90% as the first choice and 98% within the top-ten choices.

In a later work, Khorsheed [13] presented a cursive Arabic text recognition system based on HMM. This system was also segmentation-free with an easy-to-extract statistical features vector of length 60 elements, representing three different types of features. This system was trained with a data corpus which includes Arabic text of more than 600 A4-size sheets typewritten in six different computer-generated fonts: Tahoma, Simplified Arabic, Traditional Arabic, Andalus, Naskh and Thuluth. The highest achieved results were 88.7% and 92.4% for Andalus font in mono-model and tri-model, respectively. In another experiment, that system was trained with a multi-font data set that was selected randomly with same sample size from all fonts and tested with a data set consisting of 200 lines from each font, and achieved an accuracy of 95% using the tri-model.

In another effort, Krayem et al. [14] presented a word level recognition system using discrete hidden Markov classifier along with a block based discrete cosine transform. This system was trained by typewritten Arabic words in five fonts with size 14 points and lexicon size of 252 words. Vector quantization was used to map each feature vector to the closest symbol in the codebook. The multiple recognition hypotheses (N-best word lattice) of that system achieved a 97.65% accuracy. Also, the holistic approach was successfully used on the subword level. Nasrollahi and Ebrahimi [15] presented an approach to offline OCR for printed Persian subwords using wavelet packet transform. The proposed technique extracted font invariant and size invariant features from different subwords of four fonts and three sizes and compressed them using Principal Component Analysis (PCA). When tested on a subset of 2000 words of printed Persian text documents, that system achieved an accuracy of 97.9%.

In a later work [16], Slimane et al. organized the ICDAR2013 competition on multi-font and multi-size digitally represented Arabic text. The main characteristic of the winner system, Siemens system submitted by Marc-Peter Schambach et al., was the using of a three hidden layers neural network, that transforms a two-dimensional pixel plane into a sequence of class probabilities. the system have been applied on a subset of the APTI dataset [17] and managed to achieve an accuracy over 99%.

While the holistic approach avoids the challenging segmentation task of Arabic cursive scripts, it still has another challenge of dealing with large lexicon size of Arabic words. As the number of words in the lexicon grows, the recognition task becomes more computationally expensive. Most of the previously proposed holistic based Arabic OCR systems tested with small size vocabularies, but this is not practical for Arabic as a morphologically rich language with a huge vocabulary size.

In this paper, we propose a computationally efficient holistic Arabic OCR system for a large vocabulary size. For the sake of a practical approach, a lexicon reduction technique based on clustering the similar shape words is used to minimize the word recognition time. The proposed system utilizes a hybrid of several holistic features that combine global word level DCT-based features and local block based features. Using these types of features, the system manages to achieve Omni-font performance with font and size independence. Also, the presented system has a flexible architecture for integrating language modelling constraints by using a second rescoring pass for the top n-best word hypotheses. This rescoring operation provided a significant enhancement in the recognition accuracy of the system. The rest of the paper is organized as follows. Section 2 includes a description for the proposed holistic OCR system. The holistic DCT features used are described in Section 3. The developed lexicon reduction technique is illustrated in Section 4. Section 5 describes the language rescoring process used by the system. Section 6 presents system evaluation results and performance comparison with state of art commercial Arabic OCR systems. The final conclusions and prospects for future work are included in Section 7.

2. System Description

The developed holistic OCR system consists of two modules. The first one is the training module where the holistic features are extracted from the training set of the word images. The extracted features are used to build the set of clusters of similar word shapes. The generated words' clusters and their extracted features represent the knowledge base that is used in the recognition phase. The second module is the recognition module. In that module, after applying the preprocessing operations on the input image, the detected text blocks are segmented into lines and words. The features are extracted for each word image then the word cluster or best-n clusters, that have the minimum Euclidean distance with the test image vector, are assigned. The generated word list from the selected cluster is used to construct a word lattice for the possible recognition hypotheses of the whole line. This word lattice is rescored using n-gram language model to get the best recognition hypothesis. Figure 2 shows the block diagram of the proposed holistic OCR system.

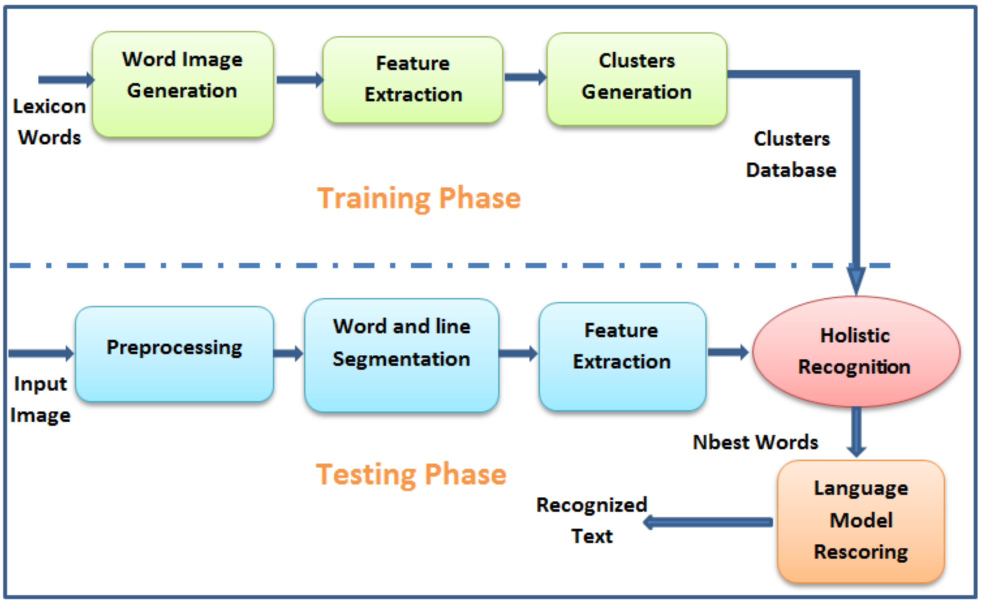

Figure 2. Block Diagram of the Holistic OCR System.

3. Feature Extraction

The main concept of the proposed algorithm is based on the property that the DCT transform compressed image is a decomposition vector which can uniquely represent the input image to be correctly reconstructed later at a decompression stage. In this work, the first 100–200 2D-DCT coefficients are used as word features that provide good approximation about the word image information. In our system, three features were experimented. Those features are: Discrete Cosine Transforms (DCT), Discrete Cosine Transforms 4-Blocks (DCT_4B), and a feature which is a combination of DCT and DCT_4B.

3.1. Discrete Cosine Transform (DCT)

The DCT features in our system are extracted via two dimensional DCT. The two dimensional DCT of an $M \times N$ image f(x, y) is defined as follows:

$$T(u,v) = \frac{1}{\sqrt{MN}} C_u C_v \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) cos(\frac{(2x+1)u\pi}{2M}) cos(\frac{(2y+1)v\pi}{2N}) \tag{1}$$

where $0 \leq x \leq M-1, 0 \leq y \leq N-1$

$$Cu = \begin{cases} \frac{1}{\sqrt{M}}, x = 0 \\ \frac{2}{\sqrt{M}}, 1 \leq x \leq M-1 \end{cases}, \ Cv = \begin{cases} \frac{1}{\sqrt{N}}, y = 0 \\ \frac{2}{\sqrt{N}}, 1 \leq y \leq N-1 \end{cases}.$$

After applying DCT to the whole word image, the features are extracted in a vector form by using the most significant DCT coefficients. The steps involved in DCT feature extraction as shown in Figure 3 are:

1. Apply the DCT to the whole word image.
2. Perform zigzag operation on the DCT coefficients I_{dct}.

The zigzag matrix I_z is a row vector matrix containing high frequency coefficients in its first N values that contain most word information. This forms features vector f_{dct} for each word.

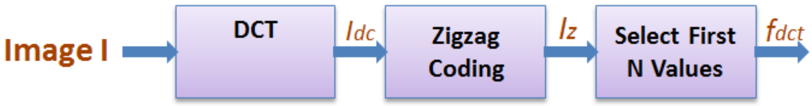

Figure 3. DCT based Feature Extraction.

3.2. Discrete Cosine Transform 4-Blocks (DCT_4B)

In this feature set, firstly we find the Centre of Gravity (COG) of image and make it as the starting point; in order to calculate the centre of gravity, the horizontal and vertical centre must be determined by the following equations:

$$C_x = \frac{M_{(1,0)}}{M_{(0,0)}} \tag{2}$$

$$C_y = \frac{M_{(0,1)}}{M_{(0,0)}} \tag{3}$$

where C_x is the horizontal centre and C_y the vertical centre of gravity and $M_{(p,q)}$ the geometrical moments of rank $p + q$:

$$M_{pq} = \sum_x \sum_y (\frac{x}{width})^p (\frac{y}{height})^q f(x,y). \tag{4}$$

The x and y determine the image word pixels. The division of x and y by the width and the height of the image, respectively, causes the geometrical moments to be normalized and be invariant to the size of the word [18]. This method uses features of COG and DCT at the same time, the first one as an auxiliary feature to divide the image into four parts and apply the second feature DCT on each part as a whole.

This feature set is extracted and implemented as follows:

1. Calculate the COG of the word image and make it as a starting point as explained in Equations (1)–(4).
2. Use the vertical and horizontal COG to divide the word image into four regions.
3. Apply the DCT to each part of the word image.
4. Perform zigzag operation on the DCT coefficients of each image part to get the first $N/4$ values that contain most word information on that word part.
5. Repeat Steps 3 and 4 sequentially for all the word parts, and then combine them together to form the feature vector of the word image.

3.3. Hybrid DCT and DCT_4B (DCT + DCT_4B)

This feature combines the two features DCT and DCT_4B.

4. Lexical Reduction and Clustering

To reduce the computation time for searching the whole lexicon in the recognition phase, the similar shape words are clustered together. The word search is performed in two steps. In the first one, the word cluster or the nearest n-clusters are determined then the best matching word inside that cluster are selected as the recognition output. For words clustering, we used the LBG algorithm [19] to cluster the words in each group depending on closeness of the word shapes from the point of view of the used features. For the clustering process, we used the same DCT and DCT_4B features that we use for the word recognition phase.

To measure the accuracy of the clustering step, and also lexical reduction, we used a clustering accuracy measure which counts the number of times the test word exists within the selected cluster/clusters per the tested words. For a vocabulary size of around 356,000 words of Simplified Arabic font (14 pt.), we tested the clustering accuracy using a test set of 3465 words and a codebook size of 1024. Table 1 shows the clustering accuracy rate of the tested words using the three implemented features when using varying number of clusters from one to 10.

Table 1. Clustering accuracy rate (percent) of Simplified Arabic font vs. number of clusters using three features (codebook size = 1024, lexicon $\simeq 356,000$).

Features	Number of Coefficients	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	Top9	Top10
DCT	160	84.7	96.0	98.4	98.9	99.1	99.4	99.5	99.6	99.7	99.7
DCT_4B	160	78.5	91.9	96.2	97.8	98.7	99.2	99.4	99.6	99.7	99.7
DCT+ DCT_4B	200	86.1	96.2	98.5	99.1	99.3	99.6	99.7	99.8	99.8	99.8

The results of Table 1 show that the DCT+DCT_4B feature is better than the other two. This hybrid feature benefited from the local and global feature of the DCT, so it achieved good results, especially in the noisy data. Figure 4 shows the relation between codebook size and clustering accuracy rate.

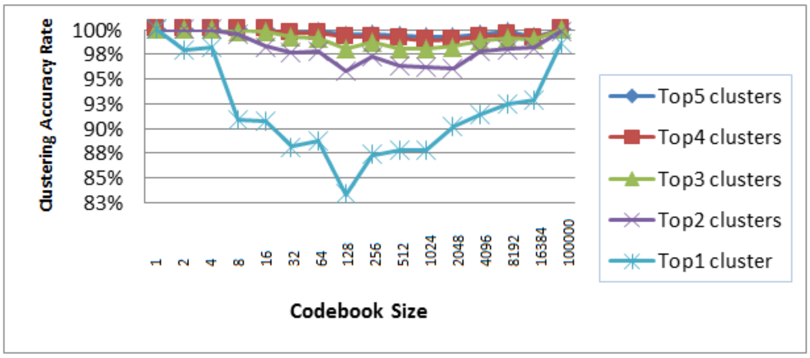

Figure 4. Clustering accuracy rate of Simplified Arabic font vs. codebook size number using DCT+ DCT_4B feature for different top clusters.

As shown in Figure 4, the clustering accuracy rate increases when using larger number of top-*n* clusters which is a logical consequence. When using a small number of clusters, each cluster contains large number of words which raises the possibility of finding the tested word within one of these clusters. When the number of clusters increase, the number of words in each cluster decrease, which reduces the clustering accuracy rate but at the same time the words within each cluster becomes more similar, which starts again to raise the clustering accuracy rate even up to the highest level when each cluster contains only one word.

5. Language Rescoring

To enhance the recognition accuracy, the top-hypotheses from the holistic recognition results are rescored using a language model. In our system, we used a 4-gram language model that was trained from a Giga-word Arabic training database [20]. The top n-hypotheses for each word are combined in a lattice format as shown in Figure 5, then we used the A* search technique to search for the best score path in that lattice using the 4-gram language model to select the best matching sentence according to the Arabic language constraints [21].

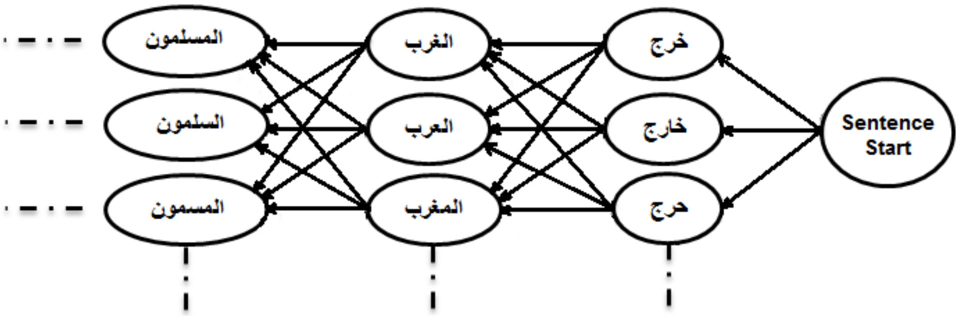

Figure 5. An example of a rescoring lattice.

6. Experiment Results

To train the proposed holistic Arabic OCR system, we used a lexicon of around 356,000 words selected from the news domain with high coverage for the Arabic Language. Using this lexicon, we generated a database of images for three fonts: Simplified Arabic, Traditional Arabic and Arabic Transparent, in 300 dpi with four different sizes.

To test the system, we used three different test datasets that represent different degrees of challenges:

1. Laser scanned text data set: This data set is composed of 1152 single words taken from newspaper articles and printed in three fonts and four different sizes in two types of qualities: clean and first copy.
2. Recent computerized books data set: A data set composed of 10 scanned pages from different recent computerized books that contain 2730 words.
3. Old un-computerized books: This data set consists of 10 scanned pages contain 2276 words from old books that are typewritten with not well known fonts.

Figure 6 illustrates some examples of the scanned images. In the first experiment, we evaluated our system using the laser scanned data set. Initially, we evaluated the system on a single font. The system was trained on a single font with single size but was tested on the same font with different sizes. We didn't use the language model with this dataset as it consists of single words. Table 2 illustrates the Word Recognition Rate (WRR) results for this experiment.

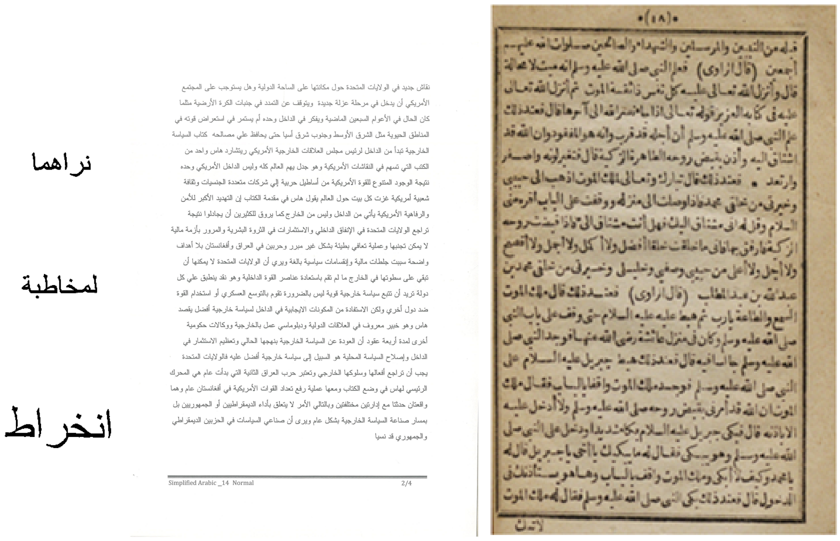
نراهما
لمخاطبة
انخراط

Figure 6. Some samples of the scanned images.

Table 2. Single font WRR (percent) for multi size fonts (12, 14, 16 and 20).

Training Font (Size)	Testing Font Name	Testing Font Size	Top 1	Top 2	Top 3	Top 4	Top 5
Simplified Arabic (14)	Simplified Arabic	12	98.26	99.48	99.91	99.96	99.96
		14	98.22	99.87	100	100	100
		16	98.39	99.44	99.78	99.83	99.83
		20	99	99.87	99.96	99.96	99.96
		Average	98.44	99.66	99.91	99.93	99.93
Arabic Transparent (14)	Arabic Transparent	12	98.13	99.61	99.96	100	100
		14	98.48	99.78	100	100	100
		16	98	99.74	99.96	100	100
		20	98.79	99.83	99.96	100	100
		Average	98.33	99.74	99.97	100	100
Traditional Arabic (16)	Traditional Arabic	12	97.57	99.65	99.83	99.96	99.96
		14	97.61	99.91	99.96	100	100
		16	97.39	99.43	99.78	99.83	99.83
		20	96.57	99.22	99.83	99.83	99.87
		Average	97.33	99.58	99.85	99.90	99.91

From the results in Table 2, we can see that the proposed system achieved very high accuracy and managed to generalize for new font sizes that were not included in the training data with best WRR of 98.44% for Simplified Arabic font and the lowest WRR of 97.33% for Traditional Arabic font. When considered the multiple recognition hypotheses, the top-5 WRR was almost 100%.

In the second experiment, the system was evaluated as omnifont by including several fonts and sizes from the laser scanned training data set. Table 3 includes the results for that evaluation.

Table 3. Multi-Fonts WRR (percent) for multi size fonts (12, 14, 16 and 20).

Training Font (Size)	Testing Font Name	Testing Font Size	Top 1	Top 2	Top 3	Top 4	Top 5
Simplified Arabic (14)	Simplified Arabic	12	98.26	99.61	100	100	100
Arabic Transparent (14)		14	98.13	99.87	100	100	100
Traditional Arabic (16)		16	98.35	99.44	99.78	99.83	99.83
		20	98.96	99.91	100	100	100
		Average	98.39	99.7	99.93	99.95	99.95
Simplified Arabic (14)	Arabic Transparent	12	98.35	99.65	99.96	100	100
Arabic Transparent (14)		14	98.96	99.87	100	100	100
Traditional Arabic (16)		16	98.74	99.83	100	100	100
		20	99.05	99.87	100	100	100
		Average	98.79	99.81	99.99	100	100
Simplified Arabic (14)	Traditional Arabic	12	97.57	99.65	99.83	99.96	99.96
Arabic Transparent (14)		14	97.61	99.91	99.96	100	100
Traditional Arabic (16)		16	97.39	99.43	99.78	99.83	99.83
		20	96.4	99.09	99.83	99.83	99.87
		Average	97.29	99.55	99.85	99.9	99.91

As we can see in Table 3, the proposed system managed to achieve for the multi-font and multi-size task almost the same WRR as the single font one. This result shows that the presented system can provide an omnifont performance.

In the third experiment, we evaluated our system using the recent and old books data sets. Table 4 shows the results of that evaluation.

Table 4. Multi-Fonts WRR (percent) for books.

Books Type	Top 1	Top 5	Top 10	Top 20
Recent Computerized	77.33	86.37	87.69	89.19
Old Uncomputerized	47.76	60.68	65.77	69.24

From the results in Table 4 we can see that our Arabic holistic OCR system achieved 77.3% WRR for recent books and 47.8% WRR for old books. Considering the top-10 hypotheses, the WRR for recent books increased to 87.7% and for old books increased to 65.7%. When considering top-20 hypotheses, the WRR increased to 89% and 69% for recent and old books, respectively. A data analysis for the recognition errors of the books data sets revealed several reasons that contributed to the reduction of the WRR. We found that this data sets included high Out Of Vocabulary (OOV) rate of around 6% for recent books and 7% for old books. It is known that the effect of the OOV is accumulative which means a single OOV word can result in recognition errors for more than one of its neighboring words. Another phenomenon that we noticed in these data set is the high rate of using the Kashida character, which was 4% for recent books and 6% for old books. The Kashida character resulted in altering the shapes of some characters which caused some word recognition errors. Also, we noticed that some fonts of the old books had large differences from the fonts used in training the system such as the Anglo-font which resulted in very low WRR for some pages.

When we applied a 4-gram language model rescoring for the books data sets using the top-10 hypothesis, we achieved 83% WRR for the recent books set and 53% WRR for the old books set. We got an absolute gain of 6% in WRR for both of the recent and old books data sets. This result show that a high percentage of the system recognition errors can be corrected using the top-n hypotheses and a language model.

In the fourth evaluation, we compared the performance of the proposed system with three commercial Arabic OCR systems, Sakhr, ABBYY and NovoDynamics, which represent the best performing Arabic OCR packages currently available. Table 5 shows these comparative results.

Table 5. Recognition rate (percent) of recent computerized and uncomputerized books. ED stands for Euclidean distance.

Books Type	NovoDynamics	Sakhr	ABBYY	Holistic (Using Top 15 with LM) Squared ED/Absolute ED
Computerized	88.45	82.17	54.33	82.97/84.76
Uncomputerized	78.15	54.94	29.22	53.21/58.04

The results in Table 5 show that, while using squared Euclidean distance as the distance measure, our system managed to achieve better performance than two systems, ABBYY and Sakhr, for the computerized books data set and achieved better performance than the ABBYY system for the uncomputerized books data set. When we used the absolute Euclidean distance, the recognition rate increased from 82.97% to 84.76% for the computerized books set and from 53.21% to 58.04% for the uncomputerized books set, and the proposed system outperformed Sakhr and ABBYY systems for both of the two datasets, although the NovoDynamics system outperfoms the proposed one. Our system is still much faster, as we will see in the next section.

As heavy computation is one of the main drawbacks for the holistic approach, we evaluated the runtime speed of the presented system. Table 6 shows the processing times of the proposed system before and after lexical reduction versus the number of selected word clusters. These experiments were run on Core i7 2.8 GHz machine with single thread execution.

Table 6. Processing time of word search and LM vs. words candidates.

Selected Words	Processing Time (s/word)
No Reduction	0.545
Lexicon Reduction (1 cluster)	0.0005
Lexicon Reduction (5 clusters)	0.0026
Lexicon Reduction (10 clusters)	0.0051

We can see from the displayed results in Table 6 that the computation cost of our developed holistic system is very practical. With lexical reduction, we managed to reduce the run time by a factor of 1000 and a one page with average number of 250 words can be computed in average time of 1.2 s compared to 1 s/page for Sakhr system, 2.3 s/page for NovoDynamics and 3.5 s/page for ABBY system.

7. Conclusions and Future Work

The holistic approaches provide effective solutions for the challenges of cursive scripts recognition such as Arabic OCR. The main drawback of such approaches is its complexity and heavy computation requirement especially for large vocabulary tasks. In this paper, we introduced a holistic Arabic OCR approach that is computationally efficient. A lexicon reduction technique based on clustering the similar shape words is utilized to reduce the word recognition time. The presented system makes use of a hybrid of several holistic features that combine global word level DCT based features and local block based features. Using this type of features, the system achieved Omni-font performance with size and font independence. Also, the suggested system has a flexible architecture to integrate language modelling constraints by using a second rescoring pass for the top n-best word hypotheses.

The proposed system has been tested using different sets of 1152 words with three different fonts and four font sizes and has achieved 99.3% WRR. It also has been tested using sets of 2730 words of recent computerized book's text and has attained more than about 84.8% WRR. Results of the holistic proposed system have been compared with known commercial Arabic OCR systems provided by the largest international and local companies, and the results were promising. In future work, we will investigate other holistic features like Wavelet Transform, Zernike Transform, Hough Transform and loci. Also, we will investigate other lexicon reduction techniques that benefit from linguistic information.

Acknowledgments: The teamwork of the "Arabic Printed OCR System" project was funded and supported by the NSTIP strategic technologies program in the Kingdom of Saudi Arabia- project no. (11-INF-1997-03). In addition, the authors acknowledge with thanks Science and Technology Unit, King Abdulaziz University for technical support.

Author Contributions: Farhan M. A. Nashwan and Mohsen A. A. Rashwan conceived and designed the experiments; Farhan M. A. Nashwan performed the experiments; Sherif M. Abdou and Mohsen A. A. Rashwan analyzed the data; Hassanin M. Al-Barhamtoshy contributed materials and analysis tools; Farhan M. A. Nashwan and Sherif M. Abdou wrote the paper; Abdullah M. Moussa substantively revised the paper.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Khorsheed, M.; Al-Omari, H. Recognizing cursive Arabic text: Using statistical features and interconnected mono-HMMs. In Proceedings of the 4th International Congress on Image and Signal Processing, Shanghai, China, 15–17 October 2011; Volume 5, pp. 1540–1543.
2. Abd, M.A.; Al Rubeaai, S.; Paschos, G. Hybrid features for an Arabic word recognition system. *Comput. Technol. Appl.* **2012** *3*, 685–691.
3. Amara, M.; Zidi, K.; Ghedira, K. An efficient and flexible Knowledge-based Arabic text segmentation approach. *Int. J. Comput. Sci. Inf. Secur.* **2017**, *15*, 25–35.
4. Radwan, M.A.; Khalil, M.I.; Abbas, H.M. Neural networks pipeline for offline machine printed Arabic OCR. *Neural Process. Lett.* **2017**, 1–19, doi:10.1007/s11063-017-9727-y.
5. El rube', I.A.; El Sonni, M.T.; Saleh, S.S. Printed Arabic sub-word recognition using moments. *World Acad. Sci. Eng. Technol.* **2010**, *4*, 610–613.
6. Madhvanath S.; Govindaraju, V. The role of holistic paradigms in handwritten word recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 149–164.

7. Falikman, M.V. Word superiority effects across the varieties of attention. *J. Rus. East Eur. Psychol* **2011**, *49*, 45–61.
8. Srimany, A.; Chowdhuri, S.D.; Bhattacharya, U.; Parui, S.K. Holistic recognition of online handwritten words based on an ensemble of svm classifiers. In Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS), Tours, France, 7–10 April 2014; pp. 86–90.
9. Erlandson, E.J.; Trenkle, J.M.; Vogt, R.C. Word-level recognition of multifont Arabic text using a feature vector matching approach. In Proceedings of the International Society for Optics and Photonics , San Jose, CA, USA, 7 March 1996; Volume 2660, pp. 147–166.
10. Al-Badr, B.; Haralick, R.M. A segmentation-free approach to text recognition with application to Arabic text. *Int. J. Doc. Anal. Recognit.* **1998**, *1*, 147–166.
11. Khorsheed, M.S.; Clocksin, W.F. Multi-font Arabic word recognition using spectral features. In Proceedings of the 15th International Conference on Pattern Recognition (ICPR-2000), Barcelona, Spain, 3–7 September 2000; pp. 543–546.
12. Khorsheed, M. A lexicon based system with multiple hmms to recognise typewritten and handwritten Arabic words. In Proceedings of the 17th National Computer Conference, Madinah, Saudi Arabia, 5–8 April 2004; pp. 613–621.
13. Khorsheed, M.S. Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK). *Pattern Recognit. Lett.* **2007**, *28*, 1563–1571.
14. Krayem, A.; Sherkat, N.; Evett, L.; Osman, T. Holistic Arabic whole word recognition using HMM and block-based DCT. In Proceedings of the 12th International Conference on Document Analysis and Recognition (DAR), Washington, DC, USA, 25–28 August 2013; pp. 1120–1124.
15. Nasrollahi, S.; Ebrahimi, A. Printed persian subword recognition using wavelet packet descriptors. *J. Eng.* **2013**, *2013*, doi:10.1155/2013/465469.
16. Slimane, F.; Kanoun, S.; El Abed, H.; Alimi, A.M.; Ingold, R.; Hennebert, J. ICDAR2013 competition on multi-font and multi-size digitally represented arabic text. In Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013; pp. 1433–1437.
17. Slimane, F.; Ingold, R.; Kanoun, S.; Alimi, A.M.; Hennebert, J. A new arabic printed text image database and evaluation protocols. In Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), Barcelona, Spain, 26–29 July 2009; pp. 946–950.
18. Zagoris, K.; Ergina, K.; Papamarkos, N. A document image retrieval system. *Eng. Appl. Artif. Intell.* **2010**, *23*, 1563–1571.
19. Linde, Y.; Buzo, A.; Gray, R. An algorithm for vector quantizer design. *IEEE Trans. Commun.* **1980**, *28*, 84–95.
20. Arabic Gigaword Fifth Edition. Available online: https://catalog.ldc.upenn.edu/LDC2003T12 (accessed on 1 March 2014).
21. Rashwan, M.A.A.; Al-Badrashiny, M.A.; Attia, M.; Abdou, S.M.; Rafea, A. A stochastic Arabic diacritizer based on a hybrid of factorized and un-factorized textual features. *IEEE Trans. Audio Speech Lang. Proc.* **2011**, *19*, 166–175.