International Journal of Engineering & Technology, 8 (1.11) (2019) 25-32

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Using Word Based Features for Word Clustering

Farhan Nashwan*[1], Mohsen Rashwan[1], Sherif Abdou[1], Hassanin Al-Barhamtoshy[2]

[1]Faculty of Engineering, Cairo University, Cairo, Egypt
[2]Faculty of Computing and Information Technology, King Abdulaziz University Jeddah, Saudi Arabia
**E-mail: hassanin@kau.edu.sa*

Abstract

In the holistic technique (HT) the whole information of the Arabic word is calculated using many possible features. In this paper, this approach is used to test the possible uses of the HT in the Arabic OCR systems. The HT is used to reduce the possible candidates for each word. We succeeded to reduce the candidates to around 115 with accuracy over 99%, given a single font and a single size from a large lexicon of more than 356K words. This vocabulary size has a good coverage for the Arabic Language. This means that the problem facing the OCR classifier is tremendously reduced, and much higher accuracy can be expected for the OCR systems.

Keywords: *holistic technique; clustering; optical character recognition; lexicon reduction*

1. Introduction

Usually in the holistic approach, a Hidden Markov Model (HMM) is used as the recognition engine. In this approach, one HMM is constructed for each distinct word in the lexicon. For this, it is inefficiently applicable to large lexicon systems due to the growing number of models in respect to the size of the lexicon. For example, for a practical problem having a lexicon size of 10,000, it is not feasible to compare with 10,000 numbers of HMM during run time against unknown inputs. That makes the problem more complicated as well as time consuming. If we look into a real-life problem, say reading amounts on bank checks or others, we see that the lexicon size is not so large. Even if the lexicon size appears larger for some practical problems, we can cope with it by reducing the lexicon size or reorganizing the search space, or using heuristics to limit the search efforts

[2]. As the number of words in the lexicon grows, the recognition task becomes more difficult and computation complexity increases. So, using lexicon reduction, the process of limiting the number of words to be compared during the recognition, can be a reasonable and powerful approach for increasing the recognition speed.

This paper presents our implemented HT for lexicon reduction. Although segmenting the words into characters performed higher recognition rates, the most difficult and important challenge is the segmentation procedure which has a direct effect on the recognition performance. The Arabic text segmentation still remains an unsolved problem though many segmentation algorithms exist. This segmentation suffers from segmentation problems such as over segmentation, under segmentation or misplaced segmentation and which affects the recognition performance in a negative way [3].

The percentage of error recognition increased in noisy text images because segmenting a word image to its letters is very difficult. One way of avoid this, is skipping segmentation and looking at words as whole entities, so-called holistic approaches.

2. Related works

Over the past several decades, many researchers have explored and developed various approaches to tackle the problem of cursiveness of Arabic script, which generally fall into two categories, holistic (global) and analytical. In the holistic approach, a word is considered as a whole unit rather than individual characters. Features are extracted from the un-segmented word and compared to a model, while in an analytical approach, a word is segmented into smaller units, which may or may not correspond to characters. . Previous research on Arabic text recognition has confirmed the difficulties in attempting to segment Arabic words into individual characters [1, 4, 5, 6].

Holistic approaches recognize an entire word directly from global shape, as a unit without segmentation, by using extraction features of the word as a whole and assign the shape or features of the word to one of the words within vocabulary by applying standard classification technique. The only drawback to this type of techniques is that they use a limited predetermined lexicon because they do not deal directly with the letters, but dealing with words can avoid this drawback by choosing a relatively large number of words covers as much as possible of the words of the Arabic language.

Copyright © 2019 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

During more than 18 years, many of the researchers have developed a number of holistic Arabic word systems; we will try to brief in this section the most important research done in this field, either with respect to handwriting or printed Arabic words. First search in this field was in 1996, Erlandson et, al. reported

a word-level recognition system for machine-printed Arabic, which computed a vector of image-morphological features (dots and hamzas, directional segments, junctions and endpoints, directional cavities, holes, descenders and intra-word gaps) on a query word image. This vector has been matched against a precomputed database of vectors from a lexicon of Arabic words in recognition stage.

In 2004, Ebrahimi and Kabir proposed a two-step method for the recognition of printed subwords, using the loci characteristic features and the k-means method, the subwords have been clustered to 300 clusters, and 10 closest clusters have been assigned. Then they used Fourier's descriptors of the subword contour to classify the input subword into the members of these 10 clusters. The training set consists of 12700 Farsi subwords in 4 different fonts and 3 sizes, and test set of 500 subwords was used. Considering the first class, top five and top ten classes, 71.4%, 95%, and 98.2% of these subwords were correctly classified. Post processing has been done using the dot features of the subwords to recognize correct subword improved the recognition rate to 92.6% [7].
In 2008, Ebrahimi and Kabir have reported on the use of characteristic loci features and k-means algorithm to cluster 113,340 printed Farsi subwords of 4 fonts and 3 sizes to 300 clusters, based on their holistic shapes. They used to test a set of 5000 subwords the clustering results were 78.71, 99.01 and 100 percent of these subwords in the first, first five and first 10 closest clusters, respectively [8].

3. Holistic technique for lexicon reduction

This part describes sets of feature extraction, clustering, grouping and clustering techniques that play a very important role in the recognition of words.

The used database in first part of work as references at training phase is prepared from two sizes of the lexicon (first one is a medium lexicon of 100K Arabic words used only for clustering method and a large lexicon of more than 356K Arabic words used for both types of techniques), which are generated by computer in Simplified Arabic, in 300 dpi and 14 size.

In the first part of the holistic technique for lexicon reduction, we present a technique to reduce the vocabulary based on clustering Arabic typewritten word images based on their holistic features. It is composed of two phases, training phase and testing phase. Figure 1 illustrates a block diagram of the proposed technique.

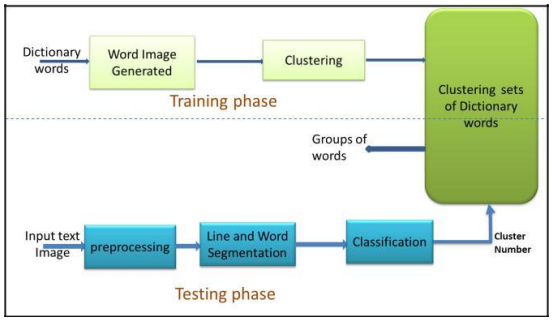

Fig. 1: Block diagram of the holistic technique using clustering

The training phase composed of image word generation, feature extraction and clustering process, where the number of clusters is up to 1024. While in the testing phase, after scanning it composed of the preprocessing, line and word segmentation, feature extraction and clustering process by using the holistic feature to select the cluster which the recognized word belongs to. We summarized the processes of training and testing phases in the following steps:
Training Phase:

- Generate the word image of all words in the dictionary.
- Extract the holistic features for all the words.
- Apply the LBG clustering for all the words to cluster each word in a cluster depending on closeness of the word shape

from the point of view of the used features.
These features are extracted simply from word images and properly categorize words into a number of clusters; the number of clusters is chosen up to 1024 as shown in Fig. 2.
Testing Phase:

- Apply preprocessing, line and word segmentation.
- Extract the holistic features of the tested image word.
- Using LBG algorithm to cluster the word in each cluster depending on closeness of the word shapes from the point of view of the used features.
- Choose closest class by doing binary search.

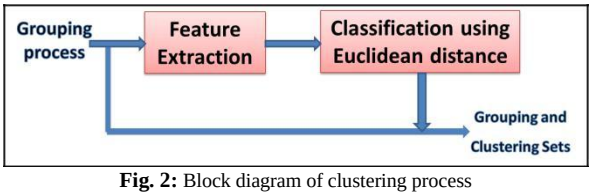

Fig. 2: Block diagram of clustering process

3.1. Feature Extraction

The proposed work has implemented many features. We have used five different features and their combinations. Those features are: Image Centroid and Cells (ICC), discrete cosine transform (DCT), dynamic time warping (DTW.), Zoning and Moments. We are going to explain these five feature sets.
In this part, we will focus on explaining in detail the features that we used in our work, whether that features we used it in data reduction or to build a proposed holistic OCR system.
The three types of features namely Image Centroid and Cell, Image Centroid and Zone and The Moment Invariant can be categorized as statistical features while the other two types of features, Discrete Cosine Transform and Dynamic Time Warping can be categorized as transformation features. On the basis of the first four types of features, we have formed the other hybrid features using different combinations of these features.
Based on their holistic features, we used all of these features to reduce the vocabulary based on grouping and clustering Arabic typewritten word images. Then we used the two features DCT and Discrete Cosine Transform 4-Blocks to build the Holistic Arabic OCR system.

Image Centroid and Cell (ICC)
Firstly we find the center of gravity (COG) of image and make it as the starting point; in order to calculate the center of gravity, the horizontal and vertical center must be determined by the
following equations:

(1,0)
=(0,0) (1)

(0,1)
horizontal center, is the vertical center of
Where is the =(0,0) (2)
gravity and moments of rank :
is the geometrical

(,) +
determine the image word pixels. The image word

The and= ∑∑ (,) (3)
binary so the is considered to be 1 when the pixel
is
and 0 when the pixel is white. The division of
is black (,)
and by the width and the height of the image, respectively,
(,)

causes the geometrical moments to be normalized and be invariant of the size of the word [9].

Secondly, we divide the image into vertical and horizontal sections (cells) of equal size (7 pixels height for each horizontal part and 8 pixels width for each vertical part), starting from the center of gravity point.

Then we form the feature set of each cell by:
1. Find the COG of image as shown in (1)-(3).
2. Use the vertical and horizontal COG to divide the image word into four sections.
3. Divide each section of the image word into vertical and horizontal cells of equal size (7 pixels height for each horizontal part and 8 pixels width for each vertical part).
4. Form the feature set of the cell by counting the following:
The number of black pixels.
The number of vertical and horizontal transitions from black to white.
5. Repeat steps 3 and 4 for all cells in 4 parts and put the values in the vector of features.
To have features of the same dimensions, though words have different dimensions, we make zero padding to the small size words.

Discrete Cosine Transform (DCT)
The DCT features in our system are extracted via two
dimensional DCT. The two dimensional DCT of an M x N
CuCv
∑X=0M-1
T(u,v)= 1 ∑y=0N-1 f(x,y) cos
image f(x, y) is defined as follows:

cos
Where (4)

0 ≤ ≤ − 1,0 ≤ ≤ − 1.
1 1
, = 0 , = 0

$= \begin{cases} \frac{1}{\sqrt{2}}, & 1 \le \le -1 \\ 1, & \end{cases} = \begin{cases} \frac{1}{\sqrt{2}}, & 1 \le \le -1 \\ 1, & \end{cases}$

After applying DCT to the whole word image, the features are extracted in a vector form by using the DCT coefficient set in a zigzag order. Usually we get the most significant DCT coefficients.

Block-Discrete Cosine Transform (BDCT)

We extract and implement this feature set in the following procedure:

1. Firstly we find the COG of the word image and make it as a start point as has already been explained in a previous ICC feature extraction method.
2. Use the COG point to divide the word image into four regions.
3. Divide each region of word image to vertical and horizontal cells of equal size (7 pixels height and 8 pixels width).
4. Apply the DCT transform for each cell $2-$ $1 \quad -1$ and
5. Get the average of the differences between all the DCT coefficients using as follows:

$$= \sum \sum ((* 2,) = (* 2 + 1,)) \quad (5)$$

$$= \sum \sum (, * 2) = (, * 2 + 1)) \quad (6)$$

$$h- \quad 1 \quad 2$$

6. Repeat steps 4 and 5 sequentially for all cells.
7. Finally, feature vector of size n (as n is the total number of cells) will be obtained for clustering. To have equal feature size through all words, zero padding is used to have equal size words.

Discrete Cosine Transform 4-Blocks(DCT-4B)

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 the 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 start point as has already been explained in a previous ICC feature extraction method in (1)-(3).

2. Use the vertical and horizontal COG to divide the word image into four regions.
3. Apply the DCT to the each part of the word image as whole.
   The zigzag matrix is a row vector matrix
4. Perform zigzag operation on the DCT coefficients . containing coefficients in its first N/4 values that high frequency contain most word information. This forms features
5. Repeat vectors for each word part.

steps 3 and 4 sequentially for all parts, then combined them together to form the feature vector of the word image.

Hybrid BDCT with Image Centroid and Cells (BDCT+ICC)

This feature combines two main features ICC and BDCT.

Hybrid DCT with Image centroid and Cells (ICC+DCT)

This feature combines two main features ICC and DCT.

Image Centroid and Zone

The steps involved in zoning feature extraction are following:

1. Firstly we find the center of gravity of the word image and make it as a start point as has already been explained in a previous ICC feature extraction method.

2. Use the COG point to divide the word image into four regions.
3. Divide each region of word image to vertical and horizontal zones of equal size (8 pixels height and 16 pixels width).
4. Compute the distance between the image centroid and each pixel present in the zone.
5. Compute the average distance between these points (in a given zone) and the centroid of the word image.
6. Repeat this procedure sequentially for all zones.
7. Finally, feature vector of size n (as n is the total number of zones) will be obtained for clustering. To have equal feature size through all words, zero padding is used to have equal size words.

Hybrid DCT and Image Centroid and Zone (DCT+ICZ)
This feature combines two main features ICZ and DCT.

Hybrid DCT and DCT-4B (DCT+ DCT-4B)
This feature combines two main features DCT and DCT-4B.

Dynamic Time Warping
Dynamic time warping (DTW) [9, 10] is an algorithm for measuring similarity between two sequences. It is a method that allows the computer to find an optimal match between two given time series. The sequences are "warped" non-linearly to determine a measure of their similarity. There are three types of features extracted from the binarized images and used in our DTW techniques. These features are Histogram, Profile and Transition extracted and implemented as follows:

1. Calculate the X-axis and Y-axis Histogram Profile by using
horizontal projection on the y-axis .
the vertical projection of the image on the x-axis , and the

2. Calculate the four contour features of the image word, which are defined as the up, down, left and right bounding contours by going along the left and the right boundaries of the image
bounding box and recording for each image column the
position of the up most and down most foreground pixels for
and lower boundaries
each image to calculate and and going along the upper
of the image's bounding box and
recording for each image row position of the left most and

the right most foreground pixels to calculate the and
left and
..This feature is calculated by going along the
right boundaries of the image bounding box and recording for each image column the number for foreground to background transitions.

3. All of obtained features are put in the feature vector to perform the clustering process using the DTW distance
measure.
To have equal feature size through all words, zero padding is used to have equal size words.

The Moment Invariant Features
Moments and functions of moments have been employed as pattern features in numerous applications to recognize two-dimensional image patterns. These pattern features extract global properties of the image such as the shape area, the center
of the mass, the moment of inertia, and so on. For a digital
is−given
order (+ −1) 1 by:
image represented in a two-dimensional array, the moment of

= ❖ ❖(,), , = 0,1,2, … . , ∞ (7)

Where M =0 =0 dimensions,
image. (,) (,) in the
and N are the horizontal and vertical
respectively, and is the intensity at point

Central moments
The first tested moments are the Central moments which are
discretized image ( ), is given (×)
invariant to scale and translation only. The two-dimensional
−1 , −1
central moments of order and for a
(−
Where = ∑ =0 ∑ =0 (−) ❖) (,) (8)
by
centroid = 00 ❖ = 00 is given by

10 and 01 are the coordinates of the
or center of gravity of the image and

(3). These central moments are invariant to the translation of the image.

Hu moments

The Hu moments are invariant under general linear transformations. However, the Hu moments are not orthogonal, so there is redundancy in the information they capture. Hu defined seven values, computed from central moments through
order three. The set of seven moment invariants of order three

1 = (20 + 02) (9)

or less as proposed by Hu is given by:

2 = (20 − 02)2 + 4 11 2 (10)

7 = (3 21 − 03)(30 + 12)[(30 + 12)2 −

3(21 0)]]− (+ 3)(+ 0)[3(+) −
+ 3 2 30 12 21 3 30 12 2
0
21 3 2

(11)

(+)

00(1+
= + 2)

Where

The seven set of moment invariants that are tested, compared and used it in this work as features as follows:

1. Firstly we find the COG of the word image and make it as a start point as has already been explained in a previous ICC feature extraction method.
2. Use the COG point to divide the word image into four regions.
3. Divide each region of word image to vertical and horizontal cells of equal size (12 pixels height and 16 pixels width).
4. Calculate the seven moment's invariant as features for each cell.
5. Repeat this procedure sequentially for all cells.
6. Finally, feature vector of size n*7 (as n is the total number of cells) will be obtained for clustering.

To have equal feature size through all words, zero padding is used to have equal size words.

3.3. LBG Clustering

Cluster analysis is the process of classifying objects into subsets that have been meaning in the context of a particular problem. Clustering process is a complementary step of the grouping process. It is done on groups that still contain a large number of words in order to distribute all the amount of vocabulary words (which in our work are more than 356K words) into groups that contain the lowest possible number of words. We used in this work a well-known algorithm, namely LBG algorithm [11]. LBG is used for clustering a set of Ns training vectors into a set of K codebook vectors. The initialization requires that the number of code words is a power of 2.

4. Experiment results

This section presents the experimental results of system that has been described in section 2. It deals with the results of the holistic technique for lexicon reduction, because the large size of the lexicon represents a major problem facing Holistic OCR system.

There are two types of results in holistic technique for lexicon reduction, the first one using clustering process only and the other using grouping and clustering.

In this part, a Dataset consists of 3465 (1155 ×3) single word images written in the same font “Simplified Arabic” in three different qualities (clean, copy1 and copy2) are tested with different features for the clustering part and the grouping and clustering. These results are based on the assumption that the test data is free of punctuation markets, and with perfect word segmentation.

At the beginning of the work, we have been used the holistic technique for lexicon data reduction using the clustering method only, but we found that the rate of the words in each cluster is still high, especially if dealing with large lexicons and needing high efficiency.

Clustering rate or accuracy of clustering, that used as measurements of the clustering process in all clustering results is defined as the rate of the number of correct tested words that exist within the selected cluster/clusters per the tested words.

Table I shows the clustering rate of data test using different features within no. of clusters when used a medium lexicon of 100K Arabic words, while table I shows the clustering rate of data test using three features within no. of clusters when used a large lexicon of 356K Arabic words. Note that the average words per cluster when we are using a medium lexicon is approximately around 100 word/cluster but when we are using a large lexicon is around 350 word/cluster.

Table 1: Clustering rate of simplified arabic font vs no. Of top clusters using different features (codebook size=1024, lexicon~100k)

Features	Coffs. NO	Top1	Top5	Top10
ICC	1716	79.0	98.4	99.4
BDCT	1144	69.9	94.4	98.0
DCT	160	88.7	99.6	99.9
DCT-4B	160	81.3	98.8	99.5

ICC+BDCT	2288	80.3	99.1	99.9
ICC+ DCT	1876	89.2	99.6	99.9
IZC	242	71.2	95.5	98.6
IZC+DCT	402	90.2	99.5	99.9
DTW	1585	92.3	99.9	99.9
Moments	1078	45.6	76.3	84.0
DCT+ DCT-4B	200	88.9	99.4	99.9

In this part of experiments, a medium size of vocabulary (about 100K words) is used with (1024 codebook size) for simplified Arabic font (14 pt.) using clustering method only.

Table I shows the clustering efficiency of the tested words using different holistic features at around 100 words /cluster multiplied with the increase in the used clusters. Two major factors play an important role in the results, distance measurements and feature type (global or local). By comparing the result in the top1 column, we can see that the global feature in DCT is better than the local feature in BDCT and DCT -4B and the DTW distance is better than the Euclidean distance.

In this part of experiments, a large size of vocabulary (about 356K words) is used with (1024 codebook size) for simplified Arabic font (14 pt.) using clustering method only. The table II shows the clustering efficiency of the tested words using three different holistic features that based on DCT feature and therefore, do not depend on the size of the font that was used in proposed Holistic OCR system at around 350 words /cluster multiplied with the increase in the used clusters.

Table2: Clustering rate of simplified arabic font vs number of clusters using three features (codebook size=1024, lexicon~356k)

Features	**Coffs. NO**	**Top1**	**Top5**	**Top10**
DCT	160	84.7	99.1	99.7
DCT-4B	160	78.5	98.7	99.7
DCT+ DCT-4B	200	86.1	99.3	99.8

The table shows the clustering rate of the tested words with three features when using multiple clusters from 1 to 10 clusters and differ from the previous table results that the Lexicon used here is about 356K words rather than 100K words, the results were limited to the features of the three most suitable for used in developing the proposal Holistic OCR system and to avoid the problem that we face in RAM if we use the features of relatively large coefficients. The results of this table show that the DCT+DCT-4B feature is better than other two because it is

composed of the hybrid of DCT and DCT-4B, and it is benefited from the local and global feature of the DCT, so it is achieved good results, especially in the noisy data. Figure 3 shows the relation between codebook size and clustering efficiency for the DCT+ DCT -4B features that used in the proposed Holistic OCR system test.

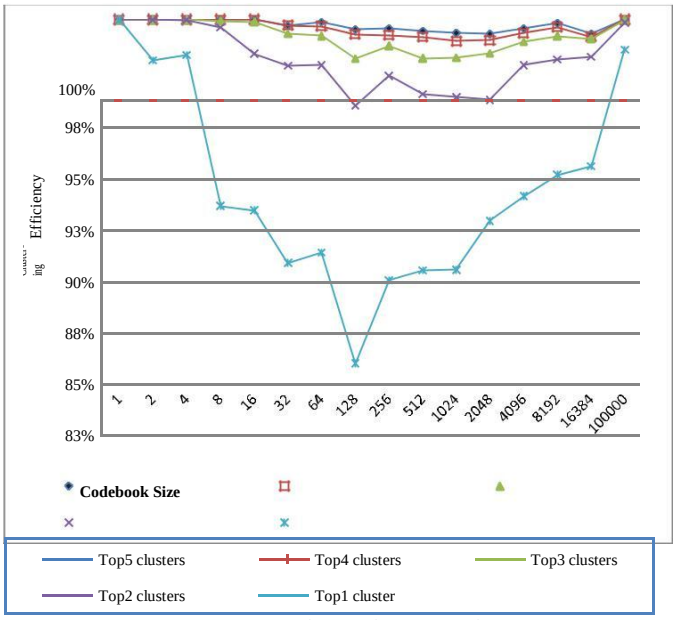

Fig. 3: Clustering rate of simplified Arabic font vs. codebook size number using DCT+ DCT-4B feature for different top clusters.

In the Fig. 3, we first show the clustering efficiency in multiple clusters used when the number of top clusters increased the clustering efficiency increased as a logical consequence.
Initially, when the number of clusters is less and therefore contain in the each cluster a large number of words, the possibilities of a tested word within one of these clusters are high so the clustering errors are minimum and then when the number of word's decrease at increasing the number of clusters the clustering efficiency decrease but at the same time the words within each cluster are more similar and thus less the percentage of clustering error and begin to improve efficiency even up to the highest level of efficiency when each cluster contains one word, and this is a special case which has been used in the proposed system.

Also through the last figure we note that the efficiency of clustering gradually decreases until it reaches the lowest value, and then resume in the increase until it reaches a value close to the value that started out. Increase the number of top clusters used greatly improves the efficiency and this is clear in the case of increasing the number of clusters from 2 onwards.

5. Results and discussion

Our target was at the beginning of the search to reach the average of 100 words per cluster and with efficiency up to 99% and the results obtained are comparable to the results of the objective of either with the average number of words per cluster or cluster efficiency.

In the clustering process, the table I shows the clustering rate of the tested words with different features when using multiple clusters from 1 to 10 clusters. For all features used, the accuracy of clustering increased with increasing the number of clusters and logical consequence of this. We note that the DTW feature is the best one; this is essentially due to the strength of the DTW distance if we compare it with Euclidean distance. But the DCT

and DCT+DCT-4B features are more acceptable for used in the proposed Holistic OCR system for two essential reasons, the first one is they are independent of the size of the fonts and the second reason is they have a less number of coefficients (160, 200 for DCT and DCT+DCT-4B respectively while in DTW 1895), which play a significant role in the search process, that represents the greatest challenge to the proposed system in addition to the fact that the Euclidean distance used is faster than DTW distance.

The table II shows the clustering rate of the tested words with three features when using multiple clusters from 1 to 10 clusters and differ from the previous table results that the lexicon used here is about 356K words rather than 100K words, the results were limited to the features of the three most suitable to be used in developing the proposed Holistic OCR system to avoid the memory leak problem if we use the features of relatively large coefficients. The results of this table show that the DCT+DCT-4B feature is better than other two ones.

6. Conclusions and future work

In this work, we presented a simple Holistic approach for typewritten Arabic OCR to capture total information for the whole Arabic word. Holistic technique is used to improve recognition accuracy and increase the speed of the process by removing unnecessary entries in the vocabulary based on grouping and clustering Arabic typewritten word images based on their holistic features. In future work, we will investigate developing the proposed holistic approach for lexicon reduction using multi-fonts and sizes rather than single font and single size. Rather than, the holistic approach will be applied in the Arabic handwritten [12] documents.

Acknowledgment

The Authors would like to thank the team of Arabic Printed OCR System" project which is supported by the NSTIP strategic technologies program in the Kingdom of Saudi Arabia- project No. (11 -INF-1997-03). Also, the authors acknowledge with thanks the Science and Technology Unit, King Abdulaziz University for the technical support. In addition, the authors thank the RDI teams who devoted no effort to help and support.

References

[1] S. Kaur, P. Mann, and S. Khurana, “Page segmentation in OCR system-A review,” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 no.3 ,pp. 420-422,2013.
[2] T. K. Bhowmik, U. Roy, and S.K. Parui, “Lexicon reduction technique for Bangla handwritten word recognition,” in Document Analysis Systems (DAS), 10th IAPR International Workshop on. IEEE, 2012.
[3] El rube', M.T.E.S., and S. S. Saleh, “Printed Arabic sub-word recognition using moments,” in World Academy of Science Engineering and Technology, vol. 4, 2010.
[4] M. Khorsheed, and H. Al-Omari, “Recognizing cursive Arabic text: using statistical features and interconnected mono-HMMs,” in Image and Signal Processing (CISP), 4th International Congress on. 2011. IEEE, vol.3, pp.1540-1543, 2011.
[5] H. Al-Barhamtoshy, and M. Rashwan, (2014). “Arabic OCR Segmented-based System”, Life Science Journal, 11 (10), (ISSN: 1097-8135), http://scholar.google.com.eg/scholar_url?hl=en&q=http://www.lifescien cesite.com/lsj/life1110/200_27304life111014_1273_1283.pdf&sa=X&scisig=AAGBfm0YM6ykkOm8jGglYVhx2mT-ZU8OIA&oi=scholaralrt, http://www.lifesciencesite.com
[6] . A. Hesham, S. Abdou, A. Badr, H. Al-Barhamtoshy, “Arabic Document Layout Analysis”, Pattern Analysis and Applications, 2017, PAAA-D-15-00373R4. http://link.springer.com/article/10.1007/s10044-017-0595-x
[7] Ebrahimi, and E. Kabir, “A two-step method for the recognition of printed subwords,” Iranian J. Electric. Comput. Eng., vol.2, no. 2, pp. 57– 62 (in Farsi), 2004.
[8] Ebrahimi, and E. Kabir, “A pictorial dictionary for printed Farsi subwords,” Pattern Recognition Letters, vol. 29, no. 5, pp. 656-663, 2008.
[9] K. Zagoris, K. Ergina, and N. Papamarkos, “A document image retrieval system,” Engineering Applications of Artificial Intelligence, vol. 23, no. 6, pp. 872-879, 2010.
[10] UniversiteitGent. “DTW algorithm,” Available at: http://www.psb.ugent.be/cbd/papers/gentxwarper/DTWalgorithm.htm, (Accessed: 21 May 2014)
[11] Myers, and L.F. HABINER, “A comparative study of several dynamic time-warping algorithms for connected-word,” Bell System Technical Journal, 1981.
[12] Abdelaziz, S. Abdou, and H. Al-Barhamtoshy, “A large vocabulary system for Arabic online handwriting recognition”, Pattern Analysis & Applications, Springer, Dec. 2015, DOI 10.1007/s10044-015-0526-7. http://link.springer.com/article/10.1007%2Fs10044-015-0526-7#page-1