Clustering of Farsi sub-word images for whole-book recognition

Article *in* Proceedings of SPIE - The International Society for Optical Engineering · February 2015

DOI: 10.1117/12.2075931

3 authors:

Mohammad Reza Soheili
Kharazmi University
8 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Ehsanollah Kabir
Tarbiat Modares University
88 PUBLICATIONS **883** CITATIONS
SEE PROFILE

Didier Stricker
Technische Universität Kaiserslautern
209 PUBLICATIONS **2,353** CITATIONS
SEE PROFILE

Clustering of Farsi Sub-word Images for Whole-book Recognition

Mohammad Reza Soheili[a,b*], Ehsanollah Kabir[a] and Didier Stricker[b]

[a] Department of Electrical and Computer Engineering, Tarbiat Modares University, Iran;
[b] German Research Center for Artificial Intelligence, Kaiserslautern, Germany

ABSTRACT

Redundancy of word and sub-word occurrences in large documents can be effectively utilized in an OCR system to improve recognition results. Most OCR systems employ language modeling techniques as a post-processing step; however these techniques do not use important pictorial information that exist in the text image. In case of large-scale recognition of degraded documents, this information is even more valuable. In our previous work, we proposed a sub-word image clustering method for the applications dealing with large printed documents. In our clustering method, the ideal case is when all equivalent sub-word images lie in one cluster. To overcome the issues of low print quality, the clustering method uses an image matching algorithm for measuring the distance between two sub-word images. The measured distance with a set of simple shape features were used to cluster all sub-word images. In this paper, we analyze the effects of adding more shape features on processing time, purity of clustering, and the final recognition rate. Previously published experiments have shown the efficiency of our method on a book. Here we present extended experimental results and evaluate our method on another book with totally different font face. Also we show that the number of the new created clusters in a page can be used as a criteria for assessing the quality of print and evaluating preprocessing phases.

Keywords: document image analysis, sub-word image, incremental clustering, shape matching, large document, Persian.

1. INTRODUCTION

Due to the rapid growth of digital libraries, digitizing large documents has become an important topic. Millions of scanned books are available in full image formats. However, efficient use of these libraries depends on the availability of robust indexing and recognition systems.

The recognition of a book is essentially different from that of a single page. This is due to the fact that a book contains much more redundant textual and pictorial information that can be used to improve the OCR performance. Another important issue is that character recognition rates as high as 95% is not yet good enough for high performance recognition of a book, since in such a large scale, as it still results in many recognition errors.

In recent years, recognition of a complete book or a document image collection has become very demanding. This is due to the fact that the performance of OCR systems can be substantially improved for large documents [1]. Hung and Hull in [2] clustered all word images of a degraded document and showed that by taking majority voting in the obtained results from an OCR software, the recognition rate can be improved significantly. Sankar and Jawahar in [3] proposed a recognition-free method for searching in collection of 500 digitized Indian books. [4] and [5] presented self-adaptable OCR systems that employ incremental learning in a recognition cycle. Rasagna et al. in [6] proposed a document-level OCR which incorporates results of a word image clustering with outputs of a normal OCR software. Kluzner et al. in [7] proposed a word image clustering-based approach for the recognition of historical texts. Sankar et al. in [8] assumed that a part of the document has been initially recognized using an OCR or human annotations. They used a hierarchical k-Means tree for the recognition of the remaining part. Xiu and Baird proposed a mutual-entropy-based model adaptation for whole book recognition [9].

* Further author information: (Send correspondence to M. R. Soheili)
M. R. Soheili : E-mail: soheili@dfki.uni-kl.de, Telephone: +49 (0)631 20 575 3510
E. Kabir.: E-mail: kabir@modares.ac.ir, Telephone: +98 (0)21 8288 3371
D. Stricker: E-mail: Didier.Stricker@dfki.de, Telephone: +49 (0)631 20 575 3500

Document Recognition and Retrieval XXII, edited by Eric K. Ringger, Bart Lamiroy
Proc. of SPIE-IS&T Electronic Imaging, Vol. 9402, 94020C © 2015 SPIE-IS&T
CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2075931

Most of the works mentioned above use character, word or sub-word image clustering for the recognition and reported promising results on English and some Indian scripts. Persian and Arabic scripts have some constraints that cause these methods not to be applicable for them. For example, these scripts have cursive style and the number and the position of dots in each character is very important. We examined a Farsi book and found out that only about five percent of its sub-words are unique. Therefore sub-word image clustering can effectively increase the speed and accuracy of the recognition of large documents, especially for old books.

Many works reported on the recognition of Arabic and Persian scripts. There are two main approaches for recognition of these scripts: segmentation-based and segmentation-free. Segmentation-based methods segment each word to containing characters or set of shape primitives and then recognize each part individually. Due to the errors occurring during segmentation, nowadays segmentation-free methods are more popular. In [10], they proposed whole sub-word recognition methods based on sub-word's holistic shape. Graves et al. in [11] described a segmentation-free method based on multidimensional recurrent neural network for recognition of Arabic handwriting. Ul-Hassan et al. in [12] and Rashid et al. in [13] used the same approach to recognize Urdu text in Nastaleeq script and low resolution Arabic printed text respectively.

In [14], we proposed a sub-word image clustering method that can be effectively used for the recognition of old printed Farsi books. In our method all sub-words, including isolated letters, even punctuation marks are clustered in an incremental manner. Unlike common methods in sub-word image clustering, which try to put sub-words with similar shapes in a cluster, our goal is to put all equivalent sub-word images in one cluster. This is not easily reachable with the low print quality. To measure the distance between sub-word images, we used an image matching algorithm based on Hamming distance and the ratio of the area to the perimeter of the connected components. Due to the fact that image matching is a time consuming process, we tried to select some features to decrease the number of image matching during clustering process, while the accuracy remain almost unchanged. For evaluation of this method, the centers of those created clusters can be labeled manually or automatically and then verified manually. This method can have lots of applications in creating digital libraries of old books.

In this paper we analyze the effects of adding more shape features on processing time, purity of clustering, and the final recognition rate. Previously published experiments have shown the efficiency of our method on a book [14]. Here we present new experimental results and evaluate our method on another book with totally different font face. We also show that the number of the new created clusters in a page can be used as a criteria for assessing the quality of print or evaluation of preprocessing phases.

The rest of this paper is organized as follows: in the next section, we review the previous related works. In section 3, we present the dataset and describe the reasons for using sub-words. We describe our previous method for sub-word image clustering in section 4. In Section 5 we present and analyze results of our experiments and also show the effect of adding more shape features on processing time and purity of clustering. We conclude the paper in section 6.

2. REVIEW RELATED WORKS

The idea of using clustering of words in optical character recognition was initially proposed by Hull et al. in [15]. They used a set of features that is referred to as the stroke direction distribution to describe the shape of words. The city-block distance was used to compare the feature vectors of two word images. They used a hierarchical, bottom-up clustering algorithm to cluster all word images and tried to define some heuristics to merge the created clusters. They showed that their method is more robust for noisy images than traditional methods.

Hung and Hull in [2] showed that about 50% of the words in a document are repeated two or more times. They improved result of a commercial OCR system from 79% to 92% with help of word image clustering. In [16] they extended their idea to detecting equivalences between portions of words and used their results for postprocessing the OCR results. Hobby and Ho in [17] proposed a character clustering method that used for enhancing degraded documents. They reported that the OCR recognition rate increased after enhancement.

Sankar and Jawahar in [3] worked on collection of 500 digitized Indian books consisting of 75,000 pages of 21 Million word images. They proposed a recognition-free method for searching in this collection which is performed on a cluster of 35 computers, for about a month. In their method all word images are clustered from the coarse to the fine level. At the finest level, the clusters would contain all instances of a given word in the collection, with all the variations in font type, style and size. At a coarse level, a large number of similar-looking words would be present in the same

cluster. They used projection profile, transition profile, Upper word profile and lower word profile as features. The features were normalized and compared using Dynamic Time Warping.

In [10] Ebrahimi and Kabir built a pictorial dictionary of 9445 sub-words that were obtained from 30000 most commonly used Farsi words. The sub-words were printed and scanned in four popular fonts and with 3 sizes. They used the characteristic loci features to cluster these sub-words. They tried to put similar sub-words in a cluster for reducing the search domain for sub-word recognition. K-Means algorithm with Euclidean distance were used for sub-word retrieval. Figure 1 shows some members of a cluster. Amount of similarity between sub-word shapes in a cluster depends on features used in clustering.

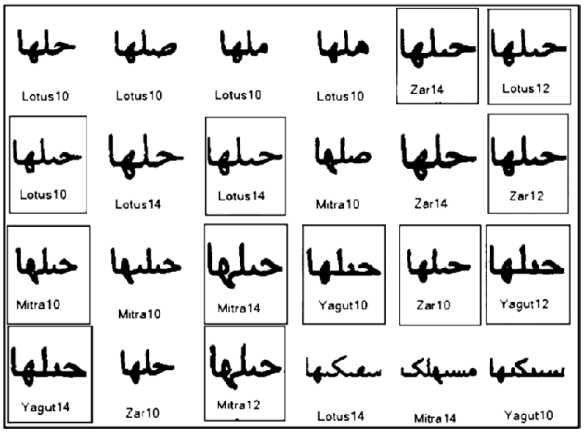


Figure 1. Some members of a cluster in the method proposed in [10]

Rasagna et al. in [6] proposed a document level OCR which incorporates information from the entire document to reduce word error rates. All word images are clustered using Locality Sensitive Hashing and also recognized using a (regular) OCR. The OCR outputs of word images in a cluster are then corrected probabilistically by comparing with the OCR outputs of other members of the same cluster.

Kluzner et al. in [7] proposed an approach to the recognition of historical texts. The system is based on clustering together all the similar words in the text. For word image comparison they used a two stages algorithm including fine and coarse, based on optical flow. All word images in each class assumed to be the same therefore the result of recognizing of them with an OCR engine also should be the same. This way, they verified and corrected the results of the OCR.

In [8] Sankar et al. assumed that a part of the document has been initially recognized using an OCR or human annotations. In the next stage a Hierarchical K-Means tree will be built from word images of this part. The word images of the document remaining part will be tested to find the nearest one in the first part. They also reported that effect of using various features including profile features, SIFT and PHOG features on accuracy of word recognition.

Lee and Smith in [18] presented a book adaptive OCR system. They used a shape clustering module in the proposed architecture to improve the recognition rate. A KD-tree-based hierarchical agglomerative clustering (KDHAC) is used in their architecture and a modified version of the template matching is used as a distance metric.

3. DATASET

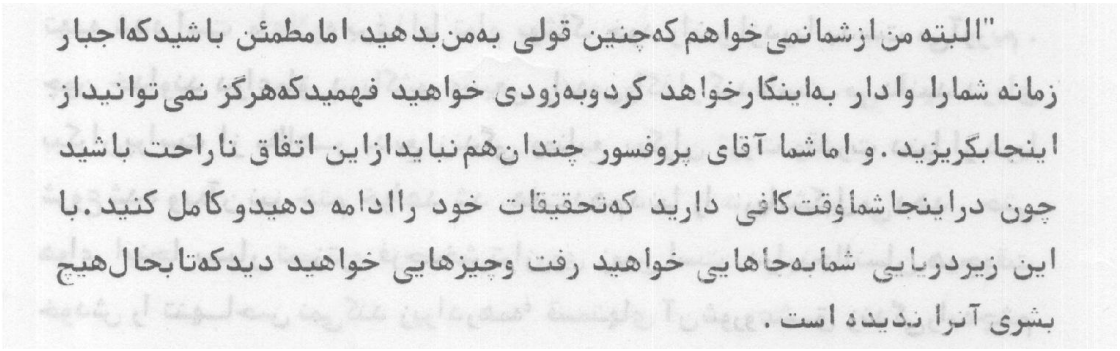
In [X] we introduce a dataset that was generated from all sub-words of a book (Book 1). The book is Farsi translation of "Twenty Thousand Leagues under the Sea" by Jules Verne, first published 46 years ago. In this paper, we add another book (Book 2) to the dataset. This book is a Farsi novel titled "There; where is not my home"† by "Reza Rahgozar", first published 32 years ago. Font face of the second book is totally different from the first one. The process of dataset

† The title is a literal translation of the original title by the authors of this paper. There is unfortunately no translation of the book in English.

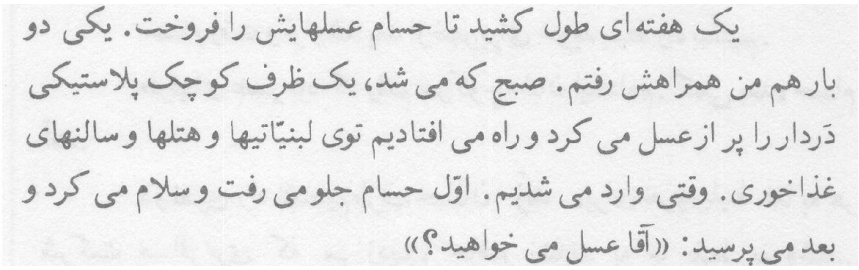
generation is the same as [14]. Table 1 shows the specification of two books. Some sample pages of these two books are available in [19]. Figure 2 shows one paragraph of each book.

Table 1. Details of the books used for the experiments

Book Title	# Pages	# Sub-words	# Unique Sub-words
Book 1	233	111576	5583
Book 2	92	47362	3200



1) Book 1



2) Book 2

Figure 2. One paragraph of each book which used for dataset generation

As mentioned in [14], we bypassed word segmentation difficulties by taking white columns as separators. We refer to the resulting segmented parts as sub-words. Due to low quality of the prints, in many cases, some parts of two adjacent words are touching each other; also some parts are overlapping horizontally. Therefore a sub-word might be a word, part of a word or multiple part of adjacent words. Figure 3, illustrates two examples of adjacent words touching each other. The ideal word segmentation and the simple sub-word segmentation are shown in this figure, too. In this paper, we do not intend to exactly segment lines to letters, words or ligatures because the print quality is too low for such accurate segmentation and the segmentation algorithm may lead to different results on the same instances of a word.

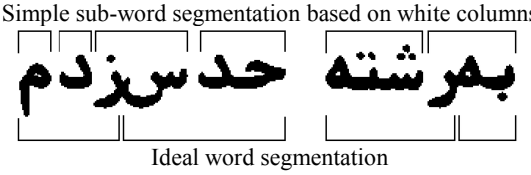


Figure 3. Two examples of word and sub-word segmentation

Figure 4 shows instances of two sub-words "به ما" and "به ر" in different pages. For both sub-words, in the first sample on the left, we do not find any segmentation point but in the last one on the right, adjacent parts can be separated easily.

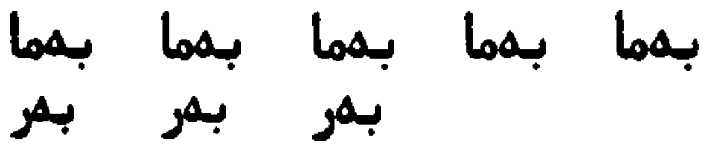


Figure 4. Different instances of two sub-words "**به ما**" and "**به ر**" in different pages

4. SUB-WORD CLUSTERING METHOD

In [14] a simple incremental algorithm was used for clustering of sub-word images. Here we briefly describe the method. For the first sample of the dataset, a cluster is created and if the next samples cannot be assigned to any existing clusters, a new cluster will be created. Each sample should be compared with centers of all existing clusters. The process of the comparison involves the following steps:

1. Four features of the sample are compared with corresponding features of the cluster center. The four features are height, width, number of black pixels and vertical distance from the baseline. The vertical distance is the distance between a sub-word bottom level and upper border of the baseline. The vertical distance is negative when a part of the sub-word is under the baseline. Figure 5 shows height, width and the vertical distance features for three sub-words of a word. To compare the features of two sub-words, the pen width is used as the threshold values for differences of height, width and the vertical distance features. The threshold value for comparing number of black pixels in two sub-words is set to one fourth of the smaller one.
2. If all feature differences are less than their corresponding thresholds, two images will be compared using an image matching algorithm. Otherwise the cluster will be removed from the candidates list. The image matching algorithm has two steps. In the first step, two images are registered and the Hamming distance between them is measured. The two images are registered by their centroids. Then we slide one image within a 5 by 5 window, pixel by pixel, around the centroid of the other one. For each position, we calculate the Hamming distance between the two images and take the position with minimum distance as the registration point. In the next step, XOR operation is performed on two registered images. Then the connected components remaining from the operation are processed. We discarded all connected components that their area is less than 16 pixels (the area of a 4 by 4 square, assuming the pen width of 8). The mentioned value is chosen to prevent the elimination of dots of letters. We decide that the two images are the same if the area is not greater than the perimeter for all of the remaining connected components. If this condition satisfies, the calculated Hamming distance is assumed as the distance of two sub-words. Otherwise we assume that two sub-word images do not belong to a cluster.
3. The proposed image matching algorithm calculates the distance between the sample and the cluster center.
4. If the two images passed the matching conditions, the distance between them will be measured and stored. Otherwise the cluster will be removed from the candidates list.

After doing above steps for all cluster centers, if no distance was saved, a new cluster is created for the sample. Otherwise the sample is assigned to the cluster that has minimum distance from its center and then the center of the selected cluster is updated. In a cluster, each member that has the minimum distance from other members is the cluster center. Those distances also are measured by the proposed image matching algorithm.

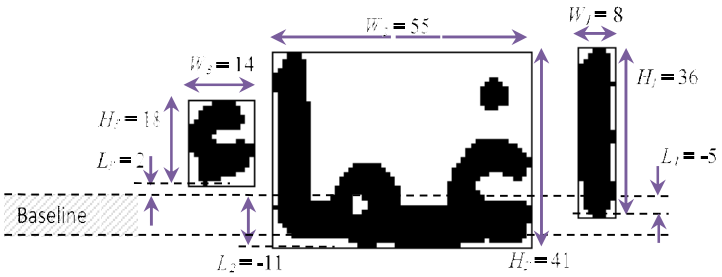


Figure 5. Three features of sub-words of a word include height(H), width(W) and the vertical distance (L)

5. EXPERIMENTAL RESULTS

In an ideal case, in Book 1 the number of clusters should be 5583 (same as the number of classes defined by the ground truth), but in our experiment 7836 clusters were formed. Vertically smeared sub-words due to the problem with paper feeder of scanner, broken and touching sub-words, non-uniformity of ink in different pages and ink spots are major causes of the increase in the number of clusters. Figure 8 shows the distribution of the samples in the clusters for both ideal and real clustering. For the ease of understanding, the clusters are divided into 10 groups based on the number of their members.

Figure 6 shows that, in an ideal case, 20 clusters have more than one thousand member and about half of the sub-words are included in these 20 clusters. However, in practice many clusters are divided into smaller ones. Ideally, the clusters of the letters " د " ," ا " and " ر " with respectively 9464, 8966 and 3425 members are the largest clusters. In practice, the clusters of the letters " ر " ," ا " ," ا "," د " and "د " with respectively 6292, 6186, 2918, 2426 and 2143 members are the largest clusters. Table 1 shows the clusters of the first group and the number of their members for both ideal (20 clusters) and real clustering (11 clusters). As shown in the table, almost half of the sub-words in a book are the most frequent letters, punctuations and conjunctions.

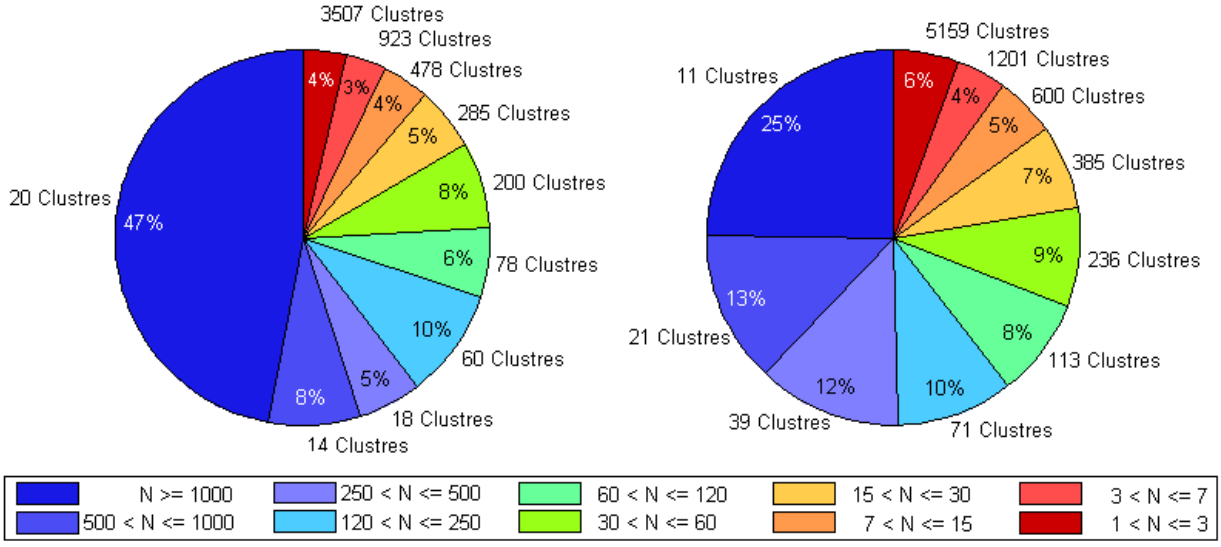


Figure 6. The distribution of the samples in the clusters in Book 1 based on the numbers of their members for both ideal (left) and real clustering (right). (N is the number of cluster members.)

Table 1. The clusters of the first group in Book 1 and the number of their members for both ideal (20 clusters) and real (11 clusters) clustering

Ideal Clustering						Real Clustering			
Cluster	No. of Members	Cluster	No. of Members	Cluster	No. of Members	Cluster	No. of Members	Cluster	No. of Members
ا	9464	ن	2236	ما	1389	د	6292	ه	1354
د	8966	را	1839	آ	1375	ا	6186	ی	1302
ر	3425	ز	1767	م	1340	ا	2918	‘ *	1126
‘ *	2948	ه	1710	ند	1232	ر	2426	.	1012
و	2921	به	1660	ین	1087	د	2143		
.	2706	که	1599	نا	1050	و	1483		
ی	2263	با	1503			و	1384		

* Single Quotation

In Book 1 only 283 of 111576 sub-words, less than 0.3 percent is assigned to wrong clusters. The main cause of the errors are eliminating of letters dots and difficulty of detection of "ک" from "گ". Some examples of occurred errors are shown in table 2. Some of these errors can be corrected in post-processing using a dictionary.

Table 2. Example of occurred errors in clustering process in Book 1

Wrong Sample	Image	بما	جر	کو	ا	کرد	سمی	مکر	نـا
	Ground Truth	بما	چر	کو	ا	کرد	سمی	مکر	نـا
Cluster Center	Image	بها	حر	کر	ا	گرد	سعی	بکر	نسا
	Ground Truth	بها	حر	کر	ا	گرد	سعی	بکر	نسا

5.1 Evaluation of the clustering results

To evaluate the clustering results, we used two criteria, purity and Rand index [20]. To compute purity, each cluster should be assigned to a class that the majority of the cluster members are belonging to. Then purity is calculated by counting the number of correctly assigned sub-words dividing by the total number of sub-words. In our experiments, the purity of 0.9975 and 0.9977 is obtained for Book 1 and 2 respectively. The purity can provide us with a preliminary estimate of recognition rate, but its disadvantage is that the number of created clusters is not considered in the calculation.

The Rand index is measured by the following equation [20]:

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{1}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. In our experiment the Rand index of 0.9897 and 0.9921 is obtained for Book 1 and 2 respectively.

5.2 Adding new features

The clustering algorithm, implemented in Matlab, takes about 23 hours to execute on a desktop computer with 64-bit WINDOWS 7 operating system, core i7 processor and 16 Gigabytes of RAM. It is too long for a book of 233 pages. More than 90% of this time is spent in image matching. Therefore we decided to include more features in order to decrease the number of performed image matching.

We added the major axis angle, MAA, and the eccentricity, ECC, as new features. MAA is the angle between the major axis of a shape and the x-axis. ECC is the measure of aspect ratio along major axis. It is the ratio of the length of major axis to the length of minor axis. MAA loses its efficiency when ECC approaches to one. Therefore, we used the combination of MAA and ECC as a feature. The selected threshold value for the MAA is 10 degrees and this feature takes into account when ECC is higher than 1.2.

The coordinates of the centroid, CC, is considered as another feature. The selected threshold value for the x and y coordinate of the centroid is half the pen width. Table 3 shows the cluster centers that should be compared with three sub-words from the first page of the book, using template matching. For each sub-word, the major axis and the centroid are displayed respectively in green and red. The table shows the effect of adding two new features to the four previous ones (height, width, relative height and number of black pixels). MAA in long sub-words tends to zero. Therefore in long sub-words CC works better than MAA, but in medium and short sub-words MAA is superior to CC because the differences between centroids are small. Overall, MAA works better since short sub-words accrued more frequently than long ones.

Table 3. Effect of adding features on the number of cluster center candidates for image matching for three sub-words with different length (green line and red dot show major axis angle and the centroid of each sub-word respectively) in Book 1.

A) Selected features: height, width, relative height and number of black pixels

B) Selected features: height, width, relative height, number of black pixels and coordinates of the centroid,

C) Selected features: height, width, relative height, number of black pixels and Pair of the major axis angle and the eccentricity

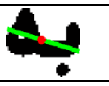
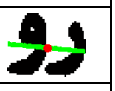
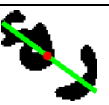
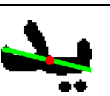
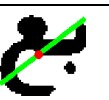
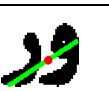
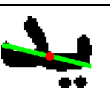
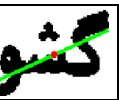
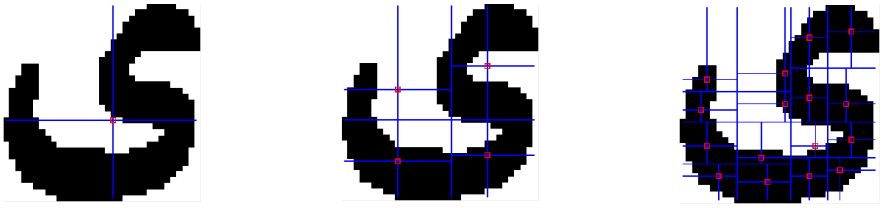
New sample	Features	Cluster center candidates for image matching						
	A							
	B	-		-		-	-	
	C	-		-	-	-	-	-
	A							
	B		-				-	-
	C	-			-	-	-	-
	A							
	B			-				-
	C		-	-		-		

Table 4 shows the effect of adding new features on Purity, Rand-Index and execution time of our algorithm for both books. Adding the coordinates of the centroid slightly changes the clustering results as the coordinates of the centroid is used in image matching process, but instead it helps us to decide about some sub-words before doing template matching, hence reducing the execution time. We obtained good results with the coordinates of the centroid, therefore we decided to investigate the effect of adding hierarchical geometric centroids proposed in [21]. In this method, the image is divided into non-overlapping regions in a recursive way. In the first step, image is divided into four region by the geometrical centroid and in the next steps, each region will be divided into four region with its own centroid. This method can adaptively partition the image to small regions based on the distribution of pixels. This feature in top levels is very robust against noise. We used the coordinates of these centroids in two first levels. Figure 7 shows the positions of the centroids in first, second and third level for "ی" character.



a b c

Figure 7. The hierarchical geometric centroids for "ی" character a) in the first level b) in the second level c) in the third level

Table 4. Effect of adding features on the number of created clusters, number of errors and clustering time.

A) Selected features: height, width, relative height and number of black pixels

B) Selected features: height, width, relative height, number of black pixels and coordinates of the centroid,

C) Selected features: height, width, relative height, number of black pixels and Pair of the major axis angle and the eccentricity

D) Selected features: height, width, relative height, number of black pixels, coordinates of the centroid and Pair of the major axis angle and the eccentricity

E) Selected features: height, width, relative height, number of black pixels and coordinates of the hierarchical geometric centroids in two levels

Book	No. Classes	Features	No. Clusters	No. Errors	Purity	Rand-Index	Time
Book 1	5583	A	7836	283	0.9975	0.9897	22h, 55m
		B	7836	283	0.9975	0.9897	9h, 58m
		C	7907	282	0.9975	0.9873	9h, 45m
		D	7907	282	0.9975	0.9873	7h, 40m
		E	8120	258	0.9977	0.9882	1h, 30m
Book 2	3200	A	3858	108	0.9977	0.9921	6h, 14m
		B	3756	103	0.9978	0.9937	3h, 8m
		C	3874	94	0.9980	0.9916	3h, 13m
		D	4000	88	0.9981	0.9909	43m
		E	4492	72	0.9985	0.9903	21m

As we can see in table 4, there is a tradeoff between number of clusters and the number of errors. However adding more features can greatly reduce the required computation time. By adding new features, number of clusters are increased but considering the speedup, it can be negligible.

5.3 Print Quality Effects

One of the most important reasons encouraging us in clustering sub-words is shown in Figure 8. Figure 8 shows the graph of proportion of new sub-words to total sub-words per pages in Book 1. One can readily observe that the number of new sub-words sharply decreases in first 30 pages in both ideal and real clustering. Between pages 30 and 100, percentage of new sub-words reduced slowly and reaches about 0.5. After page 100, new sub-words rate is almost constant. In page 58, however a sharp peak occurs, that according to our examination of this page, is due to the quality of print. In page 58, ink intensity is so low that some sub-words of the text are not readable even by human. Figure 9 shows parts of this page.

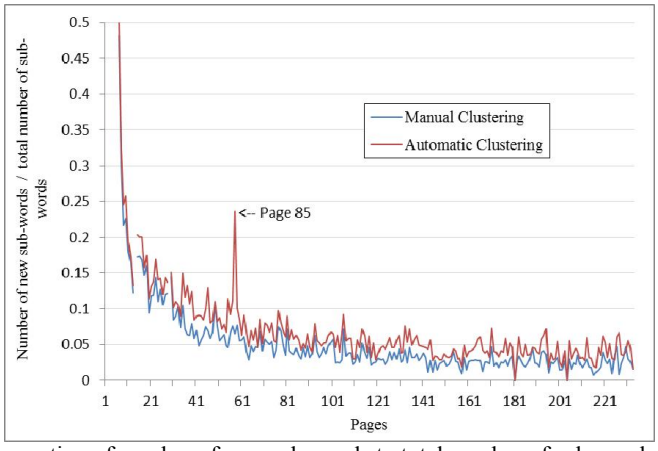


Figure 8. Graph of proportion of number of new sub-words to total number of sub-words per pages in Book 1.

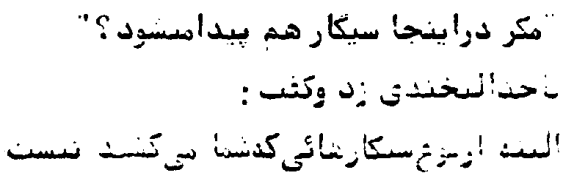


Figure 9. Part of the page 85 of Book 1 that is unreadable due to low ink

5. CONCLUSION

We described a method that exploits the redundancy of sub-words occurrences in a text, and hence making it in particular very useful for the applications dealing with large-scale documents, that might even be only available in low quality prints. One of the main advantages of our method is that it does not depend on a specific font and also it is based on a simple segmentation algorithm. Throughout the simple sub-word segmentation followed by a clustering step, assuming that all members of a cluster are the same instances of a sub-word, we can improve the performance of an OCR system. The experimental results demonstrate the efficiency of our algorithm. We achieved the purity of 0.9975 in clustering; therefore if the cluster centers are tagged manually or by a typical OCR system, we can achieve a sub-word recognition rate of 99.75 %. The experimental results, obtained for two typical books, also showed that for the recognition of all pages only less than seven percent of dictionary sub-words are needed to be recognized. We also showed that by adding some new features, we can decrease the execution time while preserving the purity of clustering and the number of clusters formed. By adding the new features to the clustering method, we could effectively reduce the clustering time more than 90%.

REFERENCES

1. K. Pramod Sankar, V. Ambati, L. Pratha, C.V. Jawahar, "Digitizing a Million Books: Challenges for Document Analysis," *Proc. of the 7th IAPR International Workshop on Document Analysis Systems* (DAS'06) 2006.
2. T. Hong, J.J. Hull, "Improving OCR performance with word image equivalence," *Symposium on Document Analysis and Information Retrieval*, V. 2, 1995.
3. K. Pramod Sankar, C.V. Jawahar, "Enabling Search Over Large Collections of Telugu Document Images – An Automatic Annotation Based Approach," *Proc. of the 5th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2006.
4. M. Meshesha, C.V. Jawahar, "Self Adaptable Recognizer for Document Image Collections," *Proc. of the 2nd international conference on Pattern Recognition and Machine Intelligence*, 2007.
5. N.V. Neeba, C.V. Jawahar, "Recognition of Books by Verification and Retraining," *Proc. of the 19th International Conf. on Pattern Recognition (ICPR 08)*, 2008.

6. V. Rasagna, A. Kumar, C.V. Jawahar, R. Manmatha, "Robust Recognition of Documents By Fusing Results of Word Clusters," *Proc. of the 10th International Conference on Document Analysis and Recognition (ICDAR 09)*, 2009.
7. V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, A. Antonacopoulos, "Word-Based Adaptive OCR for Historical Books," *Proc. of the 10th International Conference on Document Analysis and Recognition (ICDAR 09)*, 2009.
8. K. Pramod Sankar, C.V. Jawahar, R. Manmatha, "Nearest Neighbor Based Collection OCR," *Proc. of the 9th IAPR International Workshop on Document Analysis Systems (DAS'10)*, 2010.
9. P. Xiu, H.S. Baird, "Whole-Book Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, 2012.
10. A. Ebrahimi, E. Kabir, "A pictorial dictionary for printed Farsi subwords," *Pattern Recognition Letters*, vol. 29, pp. 656-663, 2008.
11. A. Graves, "Offline Arabic Handwriting Recognition with Multidimensional Neural Networks," *Book Chapter, Guide to OCR for Arabic Scripts*, Springer, 2012.
12. A. Ul-Hasan, S. Bin Ahmed, F. Rashid, F. Shafait, T.M. Breuel, "Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks," *Proc. of the 12th International Conference on Document Analysis and Recognition (ICDAR 13)*, 2013.
13. F. Rashid, M.P. Schambach, J. Rottland, S. Nüll, "Low resolution Arabic recognition with multidimensional recurrent neural networks," *Proc. of the 4th International Workshop on Multilingual OCR (MOCR '13)*, 2013.
14. M. R. Soheili, E. Kabir, D. Stricker, "Sub-word Image Clustering in Farsi Printed Books," The 7th International Conference on Machine Vision (Accepted. Draft available at http://cvs.khu.ac.ir/~soheili/ICMV2014.pdf).
15. J.J. Hull, S. Khoubyari, T.K. Ho, "World image matching as a technique for degraded text recognition," *Proc. of the 11th IAPR International Conference on Pattern Recognition Methodology and Systems*, pp. 665-668, 1992.
16. T. Hong, J.J. Hull, "Visual inter-word relations and their use in OCR postprocessing," *Proc. of the Third International Conference on Document Analysis and Recognition*, vol. 1, pp. 442-445, 1995.
17. J.D. Hobby, T.K. Ho, "Enhancing degraded document images via bitmap clustering and averaging," *Proc. of the Forth International Conference on Document Analysis and Recognition*, vol. 1, pp. 394-400, 1997.
18. D.S. Lee, R. Smith, "Improving Book OCR by Adaptive Language and Image Models," *Proc. of the 10th IAPR International Workshop on Document Analysis Systems*, 2012.
19. Online. Available at http://cvs.khu.ac.ir/~soheili/Dataset/
20. C.D. Manning, P. Raghavan, H. Schütze, "An Introduction to Information Retrieval," *Cambridge University Press*, 2009.
21. M. Yang, G. Qiu, J. Huang, and D. Elliman, "Near-Duplicate Image Recognition and Content-based Image Retrieval using Adaptive Hierarchical Geometric Centroids," Proc. of the 18th International Conference on Pattern Recognition (ICPR '06) , vol. 2, 2006.