IJNES International Journal of Natural and Engineering Sciences (6)1: 15-18, 2012
ISSN: 1307-1149, E-ISSN: 2146-0086, www.nobel.gen.tr

A Farsi/Arabic Word Spotting Approach for Printed Document Images

Yaghoub POURASAD[1] Houshang HASSIBI[2] Azam GHORBANI[3]

[1] Department of Electrical Engineering, Urmia University of Technology, Urmia, IRAN

[2] Department of Electrical and Computer Engineering, K. N. Toosi University of Technology, Tehran, IRAN

[3] Department of Engineering, Islamic Azad University, Saveh Branch, Saveh, IRAN

***Corresponding Author**
e-mail: y_pourasad@ee.kntu.ac.ir

Received : September 15, 2011
Accepted : October 21, 2011

Abstract

Word spotting is finding and locating a query word through a dataset of document images. There are many papers about English (Latin) and some papers about Arabic, but there isn't any paper about Farsi word spotting. This paper is the first paper about it. In this paper using some characteristics of Farsi scripts and some font size independent features such as number of sub words, and their aspect ratios, number of holes, dots, ascenders and descenders, and a multi level matching process, instances of a query word is found through document images. This approach has been applied on a dataset consisting of 400 Farsi document images in 4 font faces with font sizes from 8 up to 22, and precision rate 88.7% at a recall rate of 78.5% has been obtained. Proposed approach is font size independent because used features are font size independent. This approach is also applicable on Arabic and Urdu scripts.

Keywords: Farsi document image, word spotting, word searching, word image retrieval

INTRODUCTION

When we have an image from a text document, we can't search through it by common text search approaches. In this case we have two ways for searching through it. First way is OCR (Optical Character Recognition) in which, document image is converted to text and then using common text search methods, searching is done. But OCR softwares cant always transcribe document images to ASCII texts precisely. Because when document images are in degraded quality and some adjacent characters are touching each others, performance of OCR softwares falls. Also for a huge amount of document images archived in digital libraries, OCR technique requires very long time. Therefore, for retrieval purposes such as word spotting which is based on detecting and locating a few template/query images in the document image databases, resorting to a full scale OCR is wasteful and expensive. To overcome these problems researchers have proposed another way which is called keyword spotting. In keyword spotting methods, searching is done in the image domain without converting to text. In some papers such as [1] keyword spotting systems have been compared with OCR and have been shown that word spotting techniques are more efficient than OCR.

There are many statistical and structural features which are used in document image retrieval and word spotting systems, such as projection profiles, upper/lower word profiles, background to ink transitions, height, width, area, aspect ratio, moments, mean, variance, Fourier and wavelet transforms, gradient based binary features (GSC), holes, concavity, ascenders, descenders, etc. Matching methods which are used in word spotting and retrieval systems totally are two groups; training based and training free methods. Training based methods mainly are KNN (K-Nearest Neighbours), HMM (Hidden Markov Model), and BCT (B-Classification Tree) techniques. Training free methods are divided into two groups; shape code mapping and image matching approaches. In shape code mapping, a word or character is encoded into a relatively smaller set of predefined symbols, which is easier to recognize in comparison with original character set. For the retrieval of document images and word spotting, the earlier works are often based on the character shape coding that annotates characters images by a set of predefined codes [2]. The main limitation of character shape coding (CSC) techniques is their sensitivity to the character segmentation error.

To overcome the limitations of character shape coding, word shape coding schemes have been presented, in which, instead of segmenting each word image to its characters, word image itself is considered as a single component and its features without segmenting to characters are extracted. Therefore WSC (Word Shape Coding) approaches [3] are tolerant to character segmentation errors. Image to image matching are two types; the matching can be either pixel by pixel or feature based. In the pixel by pixel matching methods, template and target images are raster scanned and the distance between the corresponding pixel values is calculated. Minkowski distance measurement has been widely used for this type of matching, i.e., City Block Distance [4] and Euclidean Distance [5]. In addition to the Minkowski distance measurement techniques, some other pair wise image matching techniques have been reported. For example, in [6] (Sum of Squared Differences) SSD is used for Persian character recognition. In the feature based image to image matching, a fixed number of features are extracted and represented as feature vectors. Similarity between the template and target images is measured by comparing their corresponding feature vectors. Algorithms such as SC (Shape Context), [7], SLH [8], DTW [9, 10], and CORR [11, 12] are the most important instances of feature based image to image matching algorithms.

Most of the word spotting papers are presented for English (Latin) language and some of them are for other languages such as Chinese [5], Korean [4] Arabic [11, 13, 14], etc. There are only few papers about Arabic word spotting; and most of these papers are about handwritten Arabic documents. We haven't seen any paper about Farsi document images. In this paper an approach is presented which can be used for word spotting in Arabic, Urdu and especially Farsi printed document images.

This paper is organized as follows. In section 2 materials and methods, in section 3 results and discussion, and finally in section 4 conclusions are presented.

MATERIALS AND METHODS

There are main differences between Farsi and English scripts. therefore most of the methods which applied for English documents aren't applicable on Farsi documents. There are 32 basic characters in Farsi scripts and shape of these characters may change according to their position (beginning, middle, end or isolated) in the word. In addition, Farsi script is written from right to left and moreover, the characters of the words in Farsi texts are connected to each other both in handwritten and printed texts. In English texts, each word is composed of some letters with similar letter sizes. Also distance between characters is greater than distance between words. This feature makes it easy to segment text lines to words. But in Farsi, each word is composed of sub-words. The size of each sub-word – a part of each word that all of its letters are connected- is greatly different. Hence we use this feature for words description.

Prepprocessing

Since the document images of used database have been constructed with computer, therefore, they are noiseless and without skew; hence don't require to operations for skew correction or noise removal. Among preprocessing processes we only use two processes: text lines detection and binarization. Text line location is done by horizontal projection profile and binarization is done using Otsu's global thresholding mrthod[15].

Document images of our database were gray scale, so we had to binarize them. In order to binarize gray scale documents to binary, there are many approaches which the most common of them is Otsu's global thresholding method. In this method normalized histogram of an image is considered as a probability density function:

$$p(r_q) = \frac{n_q}{n} \qquad q=0,1,2,\dots,L-1 \tag{1}$$

Where n is total number of image pixels, n_q is number of pixels having intensity level of r_q and L is number of intensity levels in an image. In order to obtain threshold value of K, it is considered that there exist two collections: collection C_0 with level values of [0,1,2,…,K-1] and collection C_1 with level values of [K,K+1,…,L-1]. Value of K which for it inter class variance δ_B^2 be maximum, is answer as global thresholding value:

$$\delta_B^2 = \omega_0(\mu_0 - \mu_T)^2 + \omega_1(\mu_1 - \mu_T)^2 \tag{2}$$

$$\omega_0 = \sum_{q=0}^{K-1} p_q(r_q) \tag{3}$$

$$\mu_0 = \sum_{q=0}^{K-1} q p_q(r_q)/\omega_0 \tag{4}$$

$$\omega_1 = \sum_{q=K}^{L-1} p_q(r_q) \tag{5}$$

$$\mu_1 = \sum_{q=K}^{L-1} q p_q(r_q)/\omega_1 \tag{6}$$

$$\mu_T = \sum_{q=0}^{K-1} q p_q(r_q) \tag{7}$$

Feature Exraction and Searching

After binarization with threshold value of K, we extracted features from word image and searched for same features in document images. The first feature which is used to describe a word image in this method, is the number of its sub words. In order to find number of sub words in a word we can use vertical projection profile. Another feature which is used to describe a word is the aspect ratio of each sub word of that word. Aspect ratio is the width ratio over the height. In figure 1(a) a word which is composed of 6 sub words, and their bounding boxes are shown. In figure 1(b) a word which is composed of only one sub word is shown. It is clear that these two words are easily distinguishable because of their sub word numbers. As it is shown in this figure, this feature is font size independent.

Most of the Farsi characters (17 out of 32) have one, two or three dots which can be situated at the top, inside or bottom of the characters. A large number of Farsi letters (10 out of 32) have at least one hole and 6 letters out of 32 Farsi alphabet letters have ascender and 18 letters out of 32 have one or two descenders. In figure 2 the Farsi letters which have some of these features are shown.

All of these features which are used in this approach, are font size independent and are fixed for a word in all font sizes; for this reason our approach is applicable on different font sizes. In figure 3 some words and their features is illustrated.

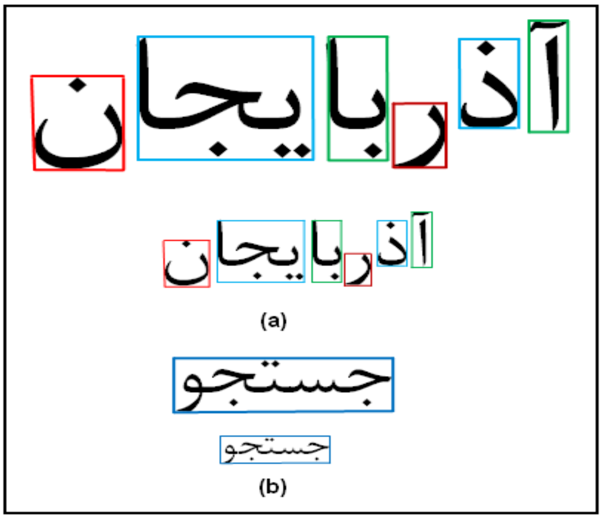
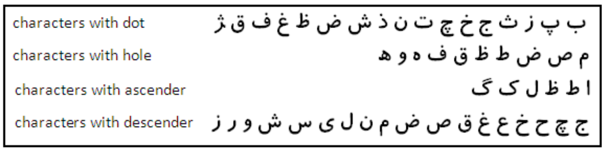


Figure1. Two words and their sub words and their bounding boxes



characters with dot	ب پ ز ث ج خ چ ت ن ذ ش ض ظ غ ف ق ژ
characters with hole	م ص ض ط ظ ق ف ه و ھ
characters with ascender	ا ط ظ ل ک گ
characters with descender	ج چ ح خ ع غ ق ص ض م ن ل ی س ش و ر ز

Figure 2. Farsi letters which have hole, dot, ascender or descender

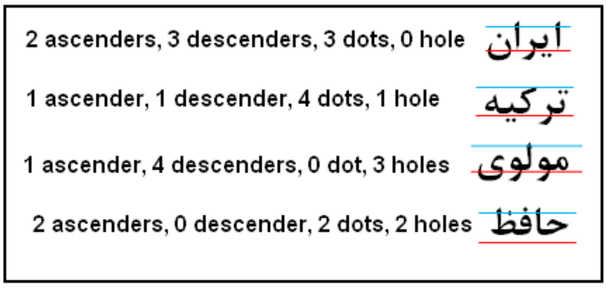

2 ascenders, 3 descenders, 3 dots, 0 hole
ایران
1 ascender, 1 descender, 4 dots, 1 hole
ترکیه
1 ascender, 4 descenders, 0 dot, 3 holes
مولوی
2 ascenders, 0 descender, 2 dots, 2 holes
حافظ

Figure 3. Some Farsi words and their numbers of ascenders, descenders, holes, and dots

RESULTS AND DISCUSSION

In our word spotting system a GUI (Graphical User Interface) has been provided so that we can enter a Farsi word by it. There are 400 computer made Farsi printed document images in our dataset in order to evaluate our method. In the used dataset, documents are in 4 different font faces e.g., 'Lotus', 'Zar', 'Nazanin', 'Mitra', and 15 font sizes from 8 up to 22.

Our spotting system has been implemented and evaluated using MATLAB software on a dual core 2.4 GHz Pentium with 512 MB RAM. Performance of a word spotting system is investigated with three criteria: (P) Precision, (R) Recall, and F1 which are defined as:

$$P = \frac{\text{No of correctly detected keywords}}{\text{No of all detected words}}$$

$$R = \frac{\text{No of correctly detected keywords}}{\text{No of actual keyword appearances}}$$

In fact precision is a criterion of correctness and recall is a criterion of completeness of searching. In some papers another evaluation criterion has been defined which indicates total performance of spotting system. This criterion is F1 and is defined as:

$$F1 = \frac{2 * P * R}{P + R}$$

Among all introduced features we first applied number of sub words and their aspect ratios. In order to compensate some font variations and also 'justifying' (Aligning text to both right and left margins) problem, we considered a tolerance value (T) for aspect ratio of sub words in each sub word. Whatever tolerance value is increased, recall rate increases but precision rate decreases and vice versa. In table1 and table2 precision and recall for six values of tolerances is presented. In table1 performance of system for different values of T (tolerance) is presented when only number of sub words and their aspect ratios are applied. In this case the best performance (best value of F1) is achieved with tolerance of 15%.

In table2 performance of system for different values of T is presented when all predefined features are used. In table2 is seen that the best performance is obtained when T= 20%.

Table 1. Performance of spotting system with different values for T, when only number of sub words and their aspect ratio features are applied

T	%5	%10	%15	%20	%25	%30
P	0.856	0.846	0.832	0.793	0.743	0.703
R	0.631	0.698	0.759	0.785	0.818	0.846
F1	0.726	0.765	0.794	0.789	0.779	0.768

Table 2. Performance of spotting system for different values for T, when all features are applied

T	%5	%10	%15	%20	%25	%30
P	0.923	0.908	0.892	0.887	0.816	0.791
R	0.631	0.698	0.759	0.785	0.818	0.846
F1	0.749	0.789	0.820	0.833	0.817	0.818

Table 3. Best precision, recall and F1 rates for some similar works and our method

paper	precision	recall	F1
Srihari [14]	55%	50%	0.524
Srihari [11]	60%	50%	0.545
Saabni [16]	85%	-	-
Our method	92.3%	84%	0.879

In fact we applied tolerance value in order to increase recall and applied large number of features in order to increase precision and therefore we achieved a high precision and high recall system. When we compare table1 and table2 results we find that using additional features leads to increasing precision rate and total performance (F1). In fact these additional features (ascenders, descenders, dots, and holes), increase P without reducing R.

In table 3 our system's performance besides other similar spotting systems is presented. As shown in table 3, our approach represents the best performance (F1); but its performance isn't perfect yet. There are some reasons for this drawback. One of them is font variations. The fact is that font face variations can change some features of a word slightly.

Another problem is related to 'justifying' (Aligning text to both the left and right margins). This problem occurs in all typing softwares such as 'Microsoft Word'. In fact these softwares while justifying a Farsi or Arabic text, extend some words along their base lines; and therefore their width and consequently their aspect ratios become different with their original's. We considered tolerance to reduce justifying and font variation problems but in order to improve system's performance we should find more robust features.

CONCLUSION

In this paper the first keyword spotting method for Farsi printed document images has been presented. In this method using number of sub words of word image, aspect ratio of their bounding boxes, number of holes, ascenders, descenders, and letters' dots, Farsi keywords are searched and found throughout

Farsi document images. This approach has been applied on a dataset consisting of 400 Farsi document images and its precision rate is 88.7% at recall rate 78.5%. Features which are used in this paper are font size independent and therefore our system is robust in font size variations. The errors which occur in this approach are because of font face variations and also justifying operation which is done when the document is written. In the future works we will try to find some features which are robust to font face variations and justifying.

REFERENCES

[1] Yue L. Tan C L. 2004. Information retrieval in document image databases. IEEE Transactions on knowledge and data engineering. 16 (11). 1398-1410.

[2] Spitz A L. 1994. Using character shape codes for word spotting in document images. Shape, Structure and pattern recognition. 22-27.

[3] Lu S, Tan C L. 2008. Retrieval of machine printed Latin documents through word shape coding. Pattern Recognition. 41. 1816-1826.

[4] Soo H K, Sang C P, Chang B J, Ji S K, Park H R, Guee S L. 2005. Keyword spotting on Korean document images by matching the keyword image. Digital Libraries, Implementing Strategies and sharing Experiences. 3815. 158-166.

[5] Yue L. Tan C L. 2004. Chinese word searching in imaged documents. International Journal of Pattern Recognition and Artificial Intelligence. 18 (2). 229-246.

[6] Pourasad Y, Hassibi H, Banaeyan M. 2011. Persian characters recognition based on spatial matching. International Review on Computers and Software. 6 (1). 55-59.

[7] Belongie S, Malik J. 2000. Matching with shape contexts. Proceedings of IEEE Workshop on Content-based access of Image and Video Libraries. 20-26.

[8] Scott G L, Languet-Higgins H C. 1991. An Algorithm for associating the features of two patterns. Proceedings of the Royal Society of London, B224. 21-26.

[9] Rath T M, Manmatha R. 2007. Word spotting for historical documents. International Journal of Document Analysis and Recognition (IJDAR). 9 (2). 139-152.

[10] Konidaris T, Gatos B, Ntzios K, Pratikakis I, Theoridis H, Perantonis S J. 2007. Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. International Journal of Document Analysis and Recognition (IJDAR). 9 (2). 166-177.

[11] Srihari S N, Srinivasan H, Huang C, Shetty S. 2006. Spotting words in Latin, Devanagari and Arabic Scripts. Indian Journal of Artificial Intelligence. 16 (3). 2-9.

[12] Zhan B, Srihari S N, Huang C. 2004. Word image retrieval using binary features", Document Recognition and Retrieval XI,SPIE. 5296. 45-53.

[13] Srihari S N, Srinivasan H, Babu P, Bhole C. 2006. Spotting words in handwritten Arabic documents. Proceedings of SPIE, San various scripts Jose, CA. 606702-1-606702-12.

[14] Srihari S N, Srinivasan H, Babu P, Bhole C. 2005. Handwritten Arabic word spotting using the CEDARABIC document analysis system. Proceedings of Symposium on Document Image Understanding Technology (SDIUT-05). 123-132.

[15] Otsu N. 1979. A threshold selection method from Gray-Level Histograms. IEEE Trans. Systems, Man. and Cybernetics, SMC-9, 62-66.

[16] Saabni R, El-Sana J. 2008. Keyword searching for Arabic Handwritten documents. Proceedings of 11th International conference on frontiers on handwritten recognition. 271-277.