Farsi/Arabic Document Image Retrieval through Sub -Letter Shape Coding for mixed Farsi/Arabic and English text

Zahra bahmani [1], Reza Azmi[2]

[1,2] Computer department
Alzahra University
Tehran, Iran

Abstract

A retrieval method for explicit recognition free Farsi/Arabic document is proposed in this paper. The system can be used in mixed Farsi/Arabic and English text. The method consists of Preprocessing, word and sub_word extraction, detection and cancelation of sub_letter connectors, annotation sub_letters by shape coding, classifier of sub_letters by use of decision tree and using of RBF neural network for sub_letter recognition. The Proposed system retrieves document images by a new sub_letter shape coding scheme in Farsi/Arabic documents. In this method document content captures through sub_letter coding of words. The decision tree-based classifier partitions the sub_letters space into a number of sub regions by splitting the sub_letter space, using one topological shape features at a time. Topological shape Features include height, width, holes, openings, valleys, jags, sub_letter ascenders/descanters. Experimental results show advantages of this method in Farsi/Arabic Document Image Retrieval.

Keywords: shape **code, sub-word, sub-letter, RBF neural network.**

1. Introduction

By Incremental use of digital libraries and the promise of paperless offices, large amount of document images are scanned and archived. Also Modern Technology has made it possible to produce, process, transmit and store digital images efficiently.

Although, the image processing technology can be used for automatic conversion of digital document images to readable text format by computer using optical recognition of characters (OCR), but this method is not practical and optimal for using in a huge volume of documents. Moreover, OCR for Persian texts still has some drawbacks. Thus new content based image retrieval techniques are required in Farsi printed documents.

Most of retrieval and recognition methods are divided into two category[1]:

The first category methods retrieval and recognation document images based on description of global shape of words or sub-words. In this method the descriptor are directly extracted from the image of the word or sub-word[1,3,4,9,10,15].

The second category methods segment a word to its letters and then extract features from the image of each letter. These features constitute a word descriptor[2,14,17].

In addition of such problems as noise, variety of fonts and size, in the segmentation based methods (especially in Farsi and cursive writing) there is the problem of segmentation. Because of the variety of shape, size and length of letters, estimation of segmentaion points of the letters is done erroneously.

In this paper we used a new segmentation method based on detection of sub-letter connectors and cancellation of them [1]. This method is free from explicit segmentation point detection. For this purpose, the connectors of main elements of the letter which have been named sub-letters, is detected. Sub-letters are exteracted by detection and cancelation of their connectors. Then sub-letters are classifier by using of their topological shape features including their height, width, holes, openings, valleys, jags, sub_letter ascenders/descanters and position of sub-letter to the base line.RBF neural network is used for recognition of sub_letters. Finally, sub-letters are encoded according to defined dictionary.

The remainder of this paper is organized as follows: secrtin2 reviews related works, section3 describe abaout Farsi script .Section 4, presents proposed system briefly. In section 5, preprocessing phase is discussed, in section 6 processing phase illustrate. Experimental results represent in section 7. Section 8 concludes the paper.

2. Review related works

In recent years, there has been much interest in the research area of Farsi Document Image and handwriting recognition [2, 3 and 4]. Some works has been done in the field of Farsi Document Image retrieval. Ebrahimi proposed a method in Persian document images recognition and retrieval. This method uses the whole shape of sub-word and sub-word's body by local features. These features are became robust against noise. Principal

Components Analysis (PCA) has been used to dimension reduction in feature space. In this proposed system, the sub-word images are clustered with k_means algorithm by using Euclidean distance for search space reduction. Two evaluation indices are used for determining the appropriate number of clusters. Moreover for clustering evaluation, the qualitative analysis has been done using classification of a test series to the centers of the clusters.

Akbari and Azmi [5] had introduced a recognition free method for Persian document images retrieval. In this method, first, upper contours of sub-words has been extracted, then a pictorial dictionary has been developed based on this featur. Document images can be retrieved by either query keywords or a query document image based on their content similarity. In another paper, Azmi and Habibi represent content based document image retrieval with support vectors clustering. In this proposed approach, sub-words feature vectors are extracted with wavelet transform and clustered with support vector clustering (SVC) algorithm.

In [6], a classification and retrival system was introducted for document images baseed on the column structure, the size of font, the density of text of regions and statistical features of continuous components of regions. This system uses this feature for document images classification and retrieval based on the visual similarity of the layout structure.

In our earlier work [21], we have proposed a Novel method for Recognition free Farsi document retrieval. In this method, the retrieval is done through recognition of sub-letters and other elements of letters such as dots and some signs like Sarkesh. Novel algorithmic technologies are used for pure Farsi/Arabic text documents. Pure text means those documents that do not include images or formulas and those in which a single language is used.

Many works had been done in English documents images retrieval [7, 8]. One of them is presented by Jilin Li, and et al [19]. They present Document Image Retrieval system with Local Feature Sequences. This paper proposed a fast, accurate and OCR-free image retrieval algorithm using local feature sequences which can describe the intrinsic, unique and page-layout-free characteristics of document images. It well handles the challenges including low resolution, different language, rotation and incompleteness and N-up.

Shijian Lu, and et al are perposed [9] a document retrieval technique that is capable of searching document images without optical character recognition (OCR). The proposed technique retrieves document images by a new word shape coding scheme, which captures the document content through annotating each word image by a word shape code. In particular, thay annotate word images by using a set of topological shape features including character ascenders/descenders, character holes, and character water reservoirs. With the annotated word shape codes, document images can be retrieved by either query keywords or a query document image.

Meshesha and Jawahar[10] described an effective word image matching scheme in their paper that achieved high performance in the presence of script variability, printing variation, degradation and word-form variants. A novel partial matching algorithm is designed for morphological matching of word form variants in a language. They formulated feature extraction scheme that extracted local features by scanning vertical strips of the word image and combining them automatically based on their discriminatory potential. They presented detailed performance analysis of the proposed approach on English, Amharic and Hindi documents.

In [11] a method was developed for automatically selecting sentences and key phrases to create a summary from an imaged document without any need for recognition of the characters in each word. In this method built word equivalence classes by using a rank blur hit-miss transform to compare word images and using a statistical classifier to determine the likelihood of each sentence being a summary sentence. Other works that used of character shape codes to document image retrieval are [12, 13, 14 and 19].

3. Farsi/Arabic Text Properties

3.1 **Farsi and Arabic Alphabets**

The Farsi alphabet has four more letters than the Arabic Alphabet which has 32 letters; whereas, the Arabic alphabet has only 28. In both Farsi and Arabic Alphabets there are several letters that share the same basic form and differ only by a small complementary part. The complementary part could be a dot, a group of dots or a slanted bar. It can lie above, below or inside the letter. In fact, all the letters in the Farsi alphabet are derived from 18 basic shapes.

3.2 **Cursiveness of the Words and Connection of the Letters**

The Arabic scripts and all of its derived forms (including Farsi Script) are inherently cursive. The letters connected to each other form a sub word. the position of each letter in the word and its preceding or following letter in the same word (if there is any), are the factors that determine the shapes of the letter. In the Farsi alphabet, similar to the Arabic alphabet, a letter can appear in four different forms according to their positions in a sub word, beginning, middle, end and single. All Farsi letters (with seven exceptions / six in Arabic) can be connected to other letters from both the right and the left sides. The seven exceptional characters can only be connected to other letters from the right side. Therefore, if any of those seven letters appear in the middle of a word, there will be a gap in connectivity. [3, 22, 21]

In the structure of the Farsi script, letters and their sub-letters are joined together by the connectors which are being along the base line. For examples, the letter "س" is written in form of " سـ "at the beginning and in the middle of the sub-word and "س" at the end of the sub-word and single state. The sub-letter for the first case consists of three jags ( ) and in the second case, it consists of two jags and one pit under the base line ( ). The letter "ی" is written in form " یـ "at the beginning and in the middle of the sub-word and "ی" at the end of the sub-word and single state. The first case has only one sub-letter that consists one jag and second case has also one sub-letter consists a hollow and one pit under base line.

4. OVERVIEW OF PROPOSED SYSTEM

Proposed system consists of 2 main phases, preprocessing and processing. Preprocessing includes binarization, text line and word extraction and overall base line detection. In processing phase each word is divided to its sub-words. Then sub-letters connectors are detected and removed from their whole body. Thus, only sub-letters remain. In the next step, Shape features of sub_letters are extracted and sub_letter annotated by its shape codes. Sub-letters space divided into a number of sub regions by use of the decision tree-based (DT) classifier and shape features. RBF are used for final recognition of sub_letters.at last sub_words are coded by their sub_letters code and this code can be used for document image content retrieval. Over view of system is shown in fig.1.

5. Pre-processing

Pre-processing phase has three steps.

Binarization: In the first step, colored and grayscale scanned document images converted to white and black ones and become binary.

Text lines and words extracted: text lines and words are extracted using blank space between lines and words. Division point between two words computed through vertical projection.

Overall base line detection: For each line, one overall base line is detected. The overall base line is detected according to this attribute that letters and their sub-letters are connected together by use of the connectors, which are being along the base line. Consequently, number of black dots along the base line is more than that of other horizontal line. Base line position is computed through horizontal projection of each line.

Noise cancelation: noises which are in the distance between lines and words are removed.

6. Processing

In this phase, sub_words are extracted from words and sub_letters are extracted from sub_words. Extracted sub_letters are given to DT. The DT classifier partitions sub_letters space to sub region by topological shape feature. Final recognition of sub_letters is done by RBF. At last all sub_words of each word are coded by use of their sub_letters cod and whole document is coded.

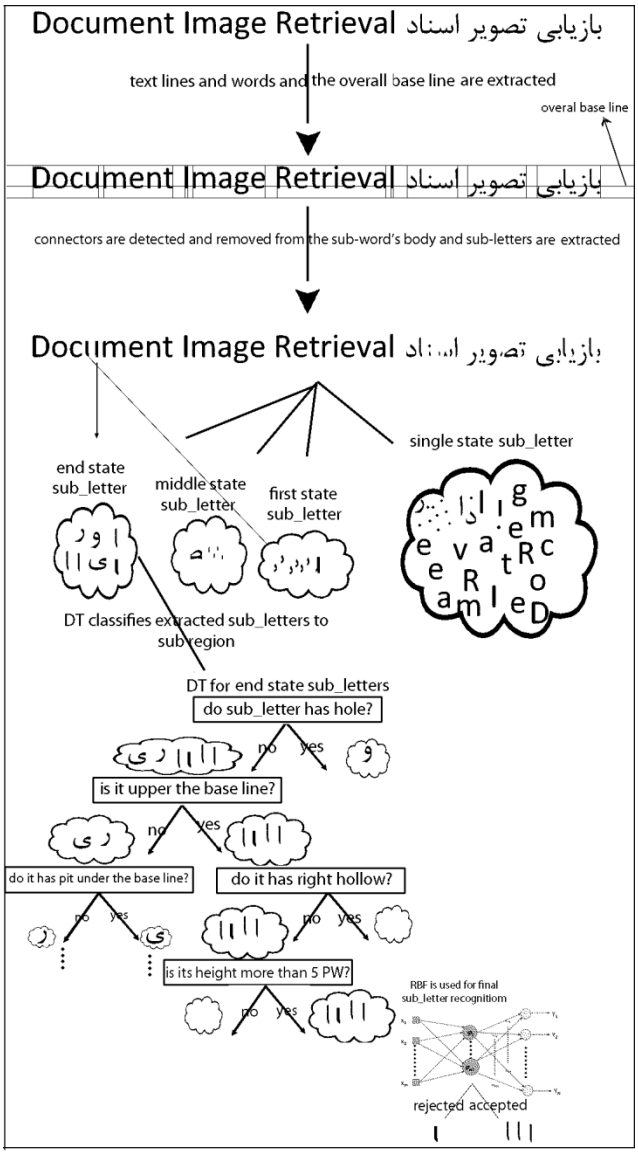

Fig. 1 over view of system

6-1. sub-letters Extraction

In considering the fact that documents will be rotated somewhat at the printing and scanning time, the overall base line usually has some error. Therefore in each word a local base line is computed through horizontal projection of word. Then, for each word, sub-words, sign and points are detected by component labeling. For sub-words processing, first connectors are detected and removed from the sub-word's body and sub-letters extracted.

6-2.topological shape feature extraction.

We used of topological shape feature for classifier of sub_letters. Features include height and width of sub_letters, holes, height and width of hole, openings, valleys, jags, sub_letter ascenders/descanters, holow and position of sub-letter to its local base line.Sub_letters are annotated by shape codes and are sended to DT classifire.

6-3.DT classifier

DT is used for classifier extracted sub_letters to sub regions. The decision tree-based classifier partitions the feature space into a number of sub regions by splitting the feature space, using one feature at a time. The regions are split until each sub region contains sub_letter that have same topological shape features. For example of single state sub_letters, letter a"ا" has only one sub_letter and its topological shape feature are height and width of sub_letters. one sub_letter is classified to sub region of "ا" if has height that is minimum six times the pen width(PW) and maximum ten and a half times PW and has width that is minimum ¾ of the PW and maximum two times the PW. According to result of experiments on mixed Farsi and English text, this error is possible occur that number "1" be classified to the class of"ا". The letter "ص" has two sub letters "ـصـ" and "ـص". The sub_letter "ـصـ" is member of middle state sub_letters and has two shape features, being upper base line (at least 2/3 of sub_letter is upper base line) and having a hole with minimum width three times of the with pen. The sub_letter "ـص" has two shape features, being under base line (at least 2/3 of sub_letter is not upper base line) and having a pit. Experimental result shows possibility of binge the sub_letter e"ی" in region of sub_lettter "ـص".

6-4. Sub-letters recognition

For final recognition of sub_letters, we used of RBF neural network and profile feature.

6.4.1 Feature extraction

Feature extraction performs in three stages

1. Image resizing: RBF recognize member of each sub region. It is fix number nodes of input layer of RBF therefore size of extracted feature vector must be fix. Then we have to do feature equalization and resize all sub_letter images of one sub region to middle size of them.

2. Profile extraction: used features are extracted profile of sub_letter. Extracted profiles for each sub_lettr image is in four directions (Up - down -left and right). These features have had reliable results in the field of printed document retrieval [23].

3. using of one dimensional Discrete Wavelet Transform (DWT): After feature extraction, the produced vector is applied to the one dimentional DWT and the output is a vector with half length of initial vector. By using DWT the system are benefited in two ways. First, it reduces the feature vector dimension. Second, it eliminates the diagram shaking caused by the jagged edge of image word during the scanning and binarization procedure.

6.4.2 NEURAL NETWORK

We used of RBF neural network in Document Image Retrieval system. The architecture of this NN is illustrated in Fig.2.

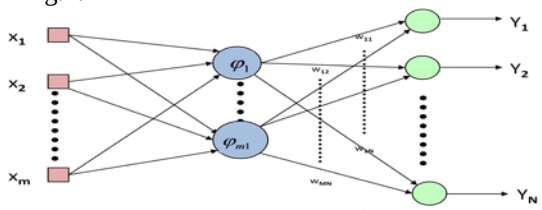

Fig. 1 Architecture of RBF NN

Architecture of RBF NEURAL NETWORK:

The RBF neural network consists of three layers, input layer, competitive layer and one output layer. Each neuron from one layer to next layer is full connected and assigned a weight to each connection.

Input layer: the extracted features are inserted into this layer.

Competitive layer: By this layer we apply a non-linear transformation from the input space to the competitive space.

Output Layer: By this layer we apply a linear transformation from the competitive space to the output space.

Training: Training is done in two phases: Unsupervised learning and supervised learning. In unsupervised learning phase the parameters of competitive layer of RBF NN are adjusted. In this stage there are 3 different strategies: first one is random selection of clusters center from data samples. Second one is use of a supervised learning method based on gradient descent that is more general case of LMS algorithm and by that minimizes total of squares of difference input and output and the third one is the using of clustering algorithms. After clustering, data are divided into special area of data space. Hence, the parameters of RBF are easily obtained from the position of centers and the distribution of clusters. Having found the parameters of RBF competitive layer, it means the number and position of centers, the weights of output layer is found by using a supervised methods like Delta rule [20].

In this research, for unsupervised learning of competitive layer, we use K_Means clustering algorithms. One supervised learning method is used to find the weights of output layer. In this, first the values of weights are set randomly and then the training samples are applied to the NN and the weights are updated according to the generated output.

Weights updating: each of the output layer units are equivalent to one of training word sets (which is one name in this case).thus, 30 training words sets produce 30 output layer units. The output value of units is computed with (1):

$$output(k) = \sum_{i=1}^{m} \varphi_{i} * w_{ik} \begin{cases} 1 & if \ \ output > 0 \\ 0 & if \ \ output \le 0 \end{cases} \quad (1)$$

Where φ_i the output of i[th] unit of competitive layer is calculated using (3) and W_{ik} is the Weight of the connection from the i[th] unit of competitive layer to the k[th] unit of output layer. Each training sample should activate its corresponding output. It means that the value of the function of the output layer unit corresponding to label of the training sample must be '1' and the values of other output layer units function must be '0'. Therefore, Weight s according to the current output and the target output will be updated by LMS rule:

$$\Delta w_i = w_i + \varepsilon \varphi_i (O_i - T_i) \quad (2).$$

Where, ε is training rate, φ_i is the output of competitive layer, O_i is the computed value of the i[th] unit, of output layer, T_i is the target output corresponding to the training sample labels. According to (2) the only weights whose corresponding outputs are different from target output are updated.

$$\varphi_i = \frac{P}{D_i + C} \quad (3).$$

Where, P and C are constant values and D_i is the distance between center of i[th] unit of competitive layer and observed sample.

Distance Function: In the proposed system, the distance between each competitive layer and training sample is computed based on a warping function.

6.5 DEFINED DICTIONARY

We defined a dictionary which includes all extracted sub-letters of sub-words. Dictionary has 4 entry classes, beginning, middle, end and single sub-letter entry. Extracted sub-letter of sub-word is shown in Tab1, 2.

7. EXPERIMENTAL RESULTS

To evaluate the proposed system we scanned 50 pages with five prevalent Farsi Font nazanin, b zar,b lotus,b mitra,b yaghot and in size 14. Percent of correct sub-letter recognition is shown in four diagrams in fig.3.

8. CONCLUSION

In this paper a retrieval method for explicit recognition free Farsi/Arabic document proposed. Proposed system retrieved the document images by a new sub_letter shape coding scheme in Farsi/Arabic documents. In the method document content captures through sub_letter coding of words. The decision tree-based classifier used for division the sub_letters space into the sub regions by splitting the sub_letter space, using topological shape features. Finally we used of RBF neural network for sub_letter recognition and coded words by their sub_word codes.

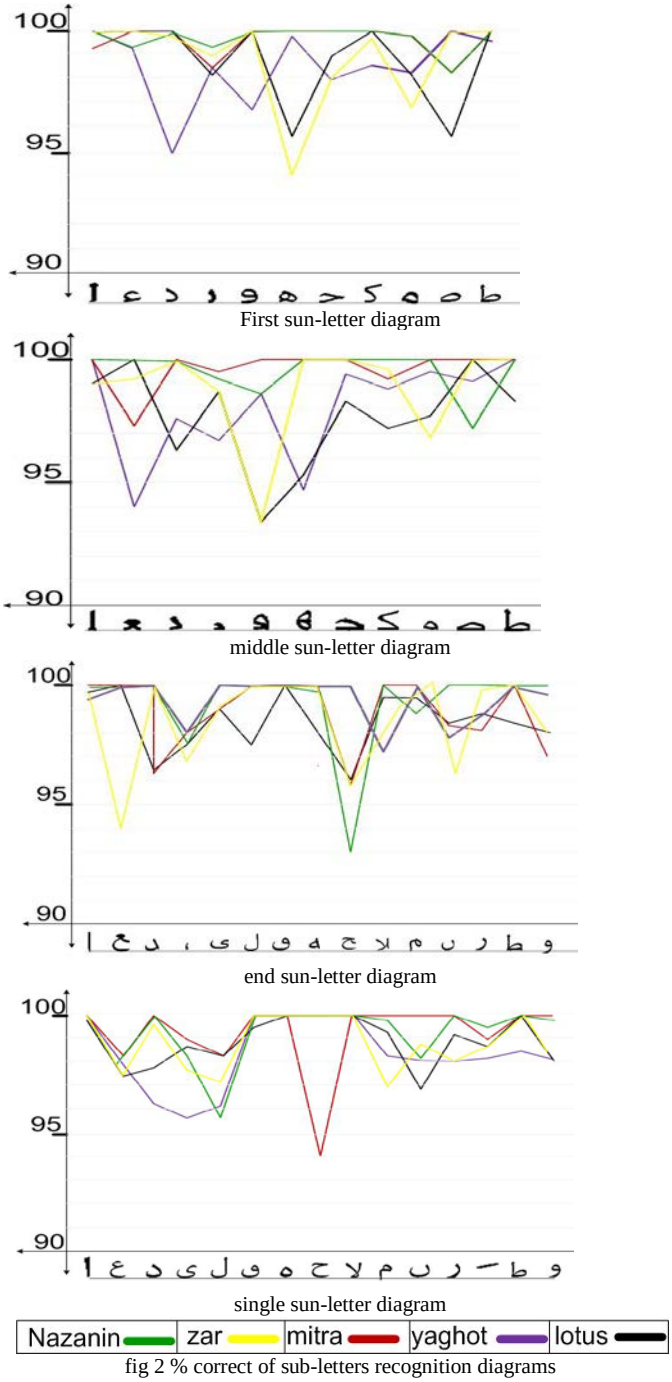

fig 2 % correct of sub-letters recognition diagrams

Table 1: letters, sub-letters and their features

num	Single	Sub Letter	Star	Sub Letter	Middle	Sub Letter	End	Sub Letter
1	آ or ا	ا	آ or ا	ا	. . . . .		ا	ا
2	ب	،،	بـ	ـ	ـبـ	ـ	ـب	،،
3	پ	،،	پـ	ـ	ـپـ	ـ	ـپ	،،
4	ت	،،	تـ	ـ	ـتـ	ـ	ـت	،،
5	ث	،،	ثـ	ـ	ـثـ	ـ	ـث	،،
6	ج	ح	جـ	ح	ـجـ	ح	ـج	ح
7	چ	ح	چـ	ح	ـچـ	ح	ـچ	ح
8	ح	ح	حـ	ح	ـحـ	ح	ـح	ح
9	خ	ح	خـ	ح	ـخـ	ح	ـخ	ح
10	د	د	.....		.....		ـد	،،
11	ذ	د	.....		.....		ـذ	،،
12	ر	ر	.....		.....		ـر	ر
13	ز	ر	.....		.....		ـز	ر
14	ژ	ر	.....		.....		ـژ	ر
15	س	ں،،	سـ	،،،	ـسـ	،،،	ـس	ں،،
16	ش	ں،،	شـ	،،،	ـشـ	،،،	ـش	ں،،

Table 2: letters, sub-letters and their features

num	Single	Sub Letter	Star	Sub Letter	Middle	Sub Letter	End	Sub Letter
17	ص	ں،	صـ	،،	ـصـ	،،	ـص	ں،
18	ض	ں،	ضـ	،،	ـضـ	،،	ـض	ں،
19	ط	ط	طـ	ط	ـطـ	ط	ـط	ط
20	ظ	ط	ظـ	ط	ـظـ	ط	ـظ	ط
21	ع	ع	عـ	ع	ـعـ	ع	ـع	ع
22	غ	ع	غـ	ع	ـغـ	ع	ـغ	ع
23	ف	و،،	فـ	و	ـفـ	ـ	ـف	،،
24	ق	ف	قـ	و	ـقـ	ـ	ـق	ف
25	ک	ک،،	کـ	ک	ـکـ	ک	ـک	ک،،
26	گ	گ،،	گـ	ک،	ـگـ	ک،	ـگ	ک،،
27	ل	ل	لـ	ا	ـلـ	ا	ـل	ل
28	م	م	مـ	ه	ـمـ	ه	ـم	م
29	ن	ں	نـ	ـ	ـنـ	ـ	ـن	ں
30	و	و					ـو	و
31	ه	ه	هـ	ه	ـهـ	ه	ـه	ه
32	ی	ی	یـ	ـ	ـیـ	ـ	ـی	ی

References

[1].Ebrahimi. A. Using printed word shape in document image retrieval and Farsi text recognition. PhD Thesis, Tarbiat Modarres University,Tehran, Iran. 2005

[2]. Hossein Khosravi ; Ehsanollah Kabir. A blackboard approach towards integrated Farsi OCR system. IJDAR, springer,2009.

[3]. Afshin Ebrahimi; Ehsanollah Kabir. A pictorial dictionary for printed Farsi sub words. Pattern Recognition Letters 29 (2008) 656–663, Elsevier, 2007.

[4]. Zahra bahmani; Reza Azmi; Fatemh Alamdar; Saman Haratizadeh. Off-Line Arabic/Farsi Handwritten Word Recognition Using RBF Neural Network and Genetic algorithm. 2th International Conference on Intelligent Computing and Intelligent Systems .ieee.2010

[5]. Reza Azmi; Mohammad Akbari; Hossain Akbari; Hossain Shirazi. LGL-DIR: Layout Graph for Layout based Document Image
Retrieval.2nd International Conference on Education Technology and Computer (ICETC). 2010, pp:262-266.

[6]. Reza Azmi; Mohammad Akbari. Document images classification and retrieval base on visual similarity. 11nd International CSI computer conference(csicc),2006.

[7] Doermann. D. The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding, 1998.

[8] . Manesh B; Kokare; M.S.Shirdhonkar. Document Image Retrieval: An Overview. International Journal of Computer Applications (0975 – 8887). 2010, Vol 1, No. 7, pp: 114-119.

[9]. Shijian L; Linlin Li; and Chew Lim Tan. Document Image Retrieval through Word Shape Coding. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 30, NO. 11, NOVEMBER 2008

[10]. Million Meshesha ;C. V. Jawahar. Matching word images for content-based retrieval from printed document images. International Institute of Information Technology, Springer-Verlag, 2008.

[11]. Chen. F. R; Bloomberg. D. S.Summarization of imaged documents without OCR. Computer Vision and Image Understanding. 1998, pp: 307–319.

[12].Smeaton A. F; Spitz A. L.Using character shape coding for information Retrieval. 4th International Conference on Document Analysis and Recognition. 1997, pp 974–978.

[13] Spitz A. L; Maghbouleh A.Text categorization using character shape Codes. SPIE Symposium on Electronic Imaging Science and Technology. 1999, pp 174–181.

[14] Spitz, A. L.Using character shape codes for word spotting in document images. Shape, Structure and Pattern Recognition. 1995, pp 382–389.

[15]. Ramin Mehran ;Hamed Pirsiavash; Farbod Razzazi. A Front-end OCR for Omni-font Persian/Arabic Cursive Printed Documents, Proceedings of the Digital Imaging Computing: Techniques and Applications (DICTA).IEEE computer society. 2005.

[16]. Arundhati Tarafdar; Ranju Mondal; Srikanta Pal; Umapada Pal Fumitaka Kimura. Shape Code based Word-image Matching for Retrieval of Indian Multi-lingual Documents. International Conference on Pattern Recognition, IEEE computer society.2010

[17]. Shahab Ensafi; Mohammad Eshghi; Mahsa Naseri. Recognition of Separate and Adjoint PersianLetters Using Primitives. IEEE Symposium on Industrial Electronics and Applications (ISIEA). October 2009.
[18]. R. Azmi; E. Kabir. A new segmentation technique for omnifont Farsi text. Pattern Recognition Letters 22 (2001) 97±104. Elsevier Science,2001
[19] T. Nakayama, "Modeling Content Identification from Document Images," Proc. Fourth Conf. Applied Natural Language Processing (ANLP '94), pp. 22-27,1994.
[20]. Mohammad S. Mohammadi,Unsupervised RBF Neural Network Training Using Genetic Algorithms,9 th Iranian Electrical Stdudent Conference , 2006.(In Farsi).
[21]. Zahra bahmani; Reza Azmi. Farsi/Arabic Document Image Retrieval Through Sub –Letter Shape Coding. To be appear, International Conference on Networks and Information ICNI Chengdu, China. November 25-27, 2011
[22]. J. Sadr; S. Izadi; F. Solimanpour. STATE-OF-THE-ART IN FARSI SCRIPT RECOGNITION Invited Paper. Center for Pattern Recognition and Machine Intelligence CENPARMI. IEEE. 2007
[23]. S.Abirami, D.Manjula, "Profile Based Information Retrieval from Printed Document Images", Computer Graphics, Imaging and Visualization (CGIV 2007).