Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 29 (2008) 656–663

Pattern Recognition Letters

www.elsevier.com/locate/patrec

A pictorial dictionary for printed Farsi subwords [☆]

Afshin Ebrahimi [a,*], Ehsanollah Kabir [b]

[a] *Department of Electrical Engineering, Sahand University of Technology (SUT), Tabriz, Iran*
[b] *Department of Electrical Engineering, Tarbiat Modarres University, Tehran, Iran*

Received 18 October 2005; received in revised form 15 October 2007
Available online 8 December 2007

Communicated by A.M. Alimi

Abstract

In this paper, we report on the use of characteristic loci features to cluster printed Farsi subwords, based on their holistic shapes. This yields a pictorial dictionary that can be used in a word recognition system to eliminate the search space. The feature vectors are compressed using PCA.

The *k*-means algorithm is used to cluster 113,340 subwords of 4 fonts and 3 sizes to 300 clusters. The minimum and maximum numbers of cluster members are 59 and 876, respectively. The mean of each cluster is used as its entry in the pictorial dictionary.

To evaluate the clustering results, a minimum mean-distance classifier was used to test a set of 5000 subwords. 78.71, 99.01 and 100 percent of these subwords were in the first, first five and first 10 closest clusters, respectively.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Pictorial dictionary; Printed words; Farsi; Persian; Arabic; Characteristic loci

1. Introduction

Shape information has important role in recognizing text images, content-based document retrieval and word spotting. Shape features of characters, subwords and words can be extracted in different ways (Mori et al., 1992; Trier et al., 1996). In order to extract pictorial information of a subword, one should extract characteristics which contain holistic information.

The use of word shapes to recognize them was proposed for the first time in (Hull, 1985). The algorithm characterizes the shape of a word by a left to right sequence of occurrences of a small number of features. This characterization is input to a classification algorithm that uses a trie, a kind of letter tree, representation of a dictionary, to locate a group or neighborhood of words that share those features. Results were reported on the predictability of the algorithm, i.e., the expected number of words in a neighborhood. The average neighborhood size of 2.5 words was reported. The application of this method to 1500 word images was resulted in 85% correct neighborhood determination.

A holistic approach to recognize handwritten English scripts has been proposed in (Powalka et al., 1997). This approach extracts ascenders and descenders of a word image and their relative locations, to describe the subwords.

A word shape recognition method for image-based retrieving of documents has been presented in (Huang et al., 2001). Document images are first segmented at word level. Local extrema points in word segments have been extracted to form vertical bar patterns. These patterns form the feature vector of a document.

A methodology for word recognition based on word shape analysis without character segmentation and recognition has been presented in (Ho et al., 1992). Global and local shape features are extracted from the image and matched with the words in the lexicon by a set of highly

[☆] This work was supported in part by the Iran Telecommunication Research Center, ITRC, under contract T500-8211.

[*] Corresponding author. Tel.: +98 412 3459374; fax: +98 412 3444322.
E-mail addresses: aebrahimi@sut.ac.ir (A. Ebrahimi), kabir@modares.ac.ir (E. Kabir).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.11.008

specialized classifiers. 94.1% correct recognition rate within the top 10 choices with a lexicon of 500 words has been reported on a set of 1671 word images.

Literature on Farsi/Arabic optical character recognition dates back to 25 years ago (Badie and Shimura, 1980; Parhami and Taraghi, 1981; Amin, 1980). Farsi and Arabic, in both printed and handwritten forms, are written cursively from right to left and segmentation of words into characters is a challenging problem (Azmi and Kabir, 2001; Zheng et al., 2004). In noisy text images, segmenting a word image to its letters is very difficult and causes errors in recognition. To reduce these errors we can employ holistic shape information. The shape of a subword can be described by contour or region descriptors (Azmi, 1999; Azmi et al., 2001; Ebrahimi, 2005).

Spectral features have been used to recognize Arabic word shapes (Khorsheed and Clocksin, 2000). Each word is transformed into a normalized polar image, and a two dimensional Fourier transform is applied to this image. The resultant spectrum tolerates variations in size, rotation or displacement. Each word is represented by a single template, and the recognition is based on the Euclidean distance from these templates. The use of hidden Markov model for Arabic word recognition is reported in (Bazzi et al., 1999; Allam, 1995; Ben Amara and Belaid, 1996).

There are few works based on holistic recognition of Farsi subwords by their shape information. Azmi has applied a pictorial dictionary for recognizing printed Farsi subwords (Azmi, 1999). He used contour features of subwords as entries of the dictionary. Each subword was represented by a sequence of labels for contour strokes, showing their position above, below or within the baseline region as well as their length as short or long. Shahreza et al. have used 45 Zernike moments as shape descriptors (Shahreza et al., 1994). They applied these descriptors on both handwritten and machine printed subwords. Ebrahimi and Kabir (2005) have proposed a two step method for the recognition of printed Farsi subwords. In the first step, each input is assigned to subword clusters by minimum mean-distance, and 10 closest clusters are found. In the second step, Fourier descriptors of the subword contour are used to classify the input subword into the members of these 10 clusters. Post-processing is done using the dots of the subwords. They reported 92.9% recognition rate.

In this paper we use a region-based method to represent the holistic shapes of Farsi subwords and build a pictorial dictionary. The characteristic loci features of the main part of each subword, after removing its dots, are found. Feature reduction is done by principle component analysis. The subwords are clustered by k-means algorithm. The mean of each cluster represents an entry to the dictionary.

The paper is organized as follows. In the next section, some characteristics of Farsi script are presented. Dataset generation is explained in Section 3. Different steps in building the pictorial dictionary are explained in Section 4. Section 5 describes the use of this dictionary in subword recognition. Section 6 concludes the paper.

2. Farsi script

Farsi is written from right to left with 32 letters. Some Farsi letters like “ب”, “پ”, “ت”, and “ث”, have similar bodies and their only differences are in the number and positions of their dots. The letters connected to each other form a subword. For example, “مهـــر” is a subword formed by three letters, “مـ”, “ـهـ” and “ـر”. One or more subwords could form a word. The word “مهـرگـــان” is formed by three subwords, “مهـــر”, “گـــا” and “ن”. Shape of letters may vary according to their positions in a subword, i.e. beginning, middle and end (Table 1). For example, the letters “ف” and “ق” have different bodies when used as a single letter or at the end of a subword, but they have similar bodies when come in the middle or at the beginning of a subword. There are 7 letters, “ا”, “د”, “ذ”, “ر”, “ز”, “ژ” and “و” which do not connect to the next letter. The neighboring characters, separated or connected, may overlap vertically. These characteristics of Farsi script are shown in Fig. 1 (Azmi and Kabir, 2001).

3. Dataset generation

The electronic databases of some Farsi newspapers were used to collect a set of Farsi words (Razavi and Kabir, 2004). 30,000 most commonly used words were selected. 12,700 subwords were obtained from these words. There are many subwords whose differences are only in their dots. Removing the dots reduces the number of subwords to 9445. The distribution of subwords and their bodies in our database is shown in Fig. 2.

The biggest reduction in the number of subwords by removing their dots occurred on subwords of 3 letters. As shown in Fig. 2, total number of subwords of 3 letters in our dataset is 3073. Since some subword bodies are the same, if the dots of the subwords are removed, the number of subword bodies of 3 letters is reduced to 1606.

The subword bodies were printed by a laser printer in four fonts, Lotus, Mitra, Yagut and Zar, with 3 sizes 10, 12 and 14, and scanned in 400 dpi. Therefore, our dataset consists of a total number of 113340 subword bodies (Fig. 3). The reason for using these fonts is their popularity in Farsi books, magazines, newspapers and official documents.

4. Building the pictorial dictionary

To recognize a subword based on its holistic shape, it is required to compare it with all subwords in the database. By using a pictorial dictionary, one can partition the search space, thus reducing the complexity of the recognition process.

Table 1
Different forms of Farsi letters depending on their position in a subword

	Single	Start	Middle	End		Single	Start	Middle	End
1	ا or آ	ا or آ	---	ـا	17	ص	صـ	ـصـ	ـص
2	ب	بـ	ـبـ	ـب	18	ض	ضـ	ـضـ	ـض
3	پ	پـ	ـپـ	ـپ	19	ط	طـ	ـطـ	ـط
4	ت	تـ	ـتـ	ـت	20	ظ	ظـ	ـظـ	ـظ
5	ث	ثـ	ـثـ	ـث	21	ع	عـ	ـعـ	ـع
6	ج	جـ	ـجـ	ـج	22	غ	غـ	ـغـ	ـغ
7	چ	چـ	ـچـ	ـچ	23	ف	فـ	ـفـ	ـف
8	ح	حـ	ـحـ	ـح	24	ق	قـ	ـقـ	ـق
9	خ	خـ	ـخـ	ـخ	25	ك	کـ	ـکـ	ـك
10	د	---	---	ـد	26	گ	گـ	ـگـ	ـگ
11	ذ	---	---	ـذ	27	ل	لـ	ـلـ	ـل
12	ر	---	---	ـر	28	م	مـ	ـمـ	ـم
13	ز	---	---	ـز	29	ن	نـ	ـنـ	ـن
14	ژ	---	---	ـژ	30	و	---	---	ـو
15	س	سـ	ـسـ	ـس	31	ه	هـ	ـهـ	ـه
16	ش	شـ	ـشـ	ـش	32	ي	يـ	ـيـ	ـي

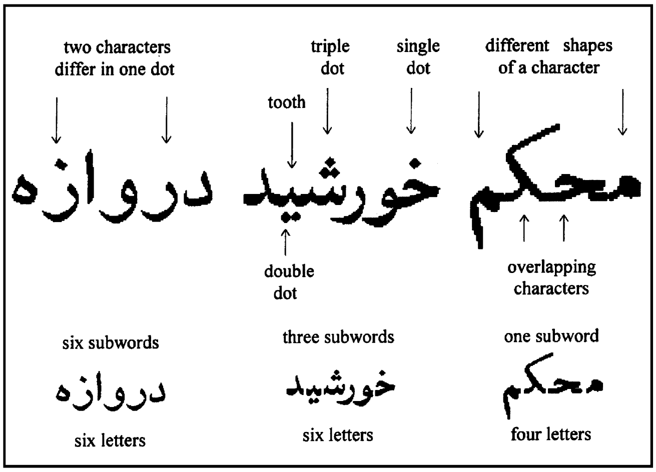


Fig. 1. Some features of Farsi script (Azmi and Kabir, 2001).

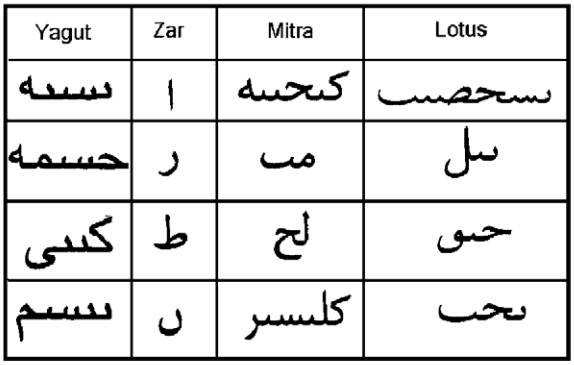

Yagut
Zar
Mitra
Lotus

Fig. 3. Samples of Farsi subword bodies.

The block diagram of our method to build a dictionary is presented in Fig. 4. We use characteristic loci features to represent the subwords. The feature vectors are compressed by PCA, reducing the number of features from 256 to 27.

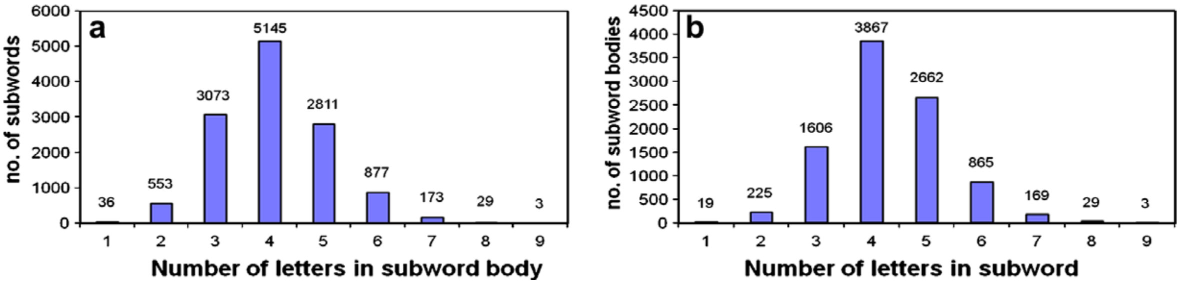


Fig. 2. Distribution of subwords (a) and subword bodies (b) vs. number of their letters.

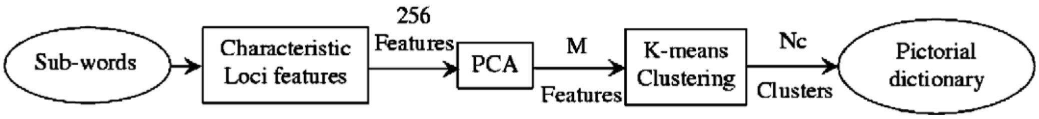


Fig. 4. Block diagram of building the pictorial dictionary.

Table 2
Correct classification rate of subwords in cluster level to the first, first 5 and first 10 closest clusters

Feature	Number of features	First cluster (%)	Top 5 clusters (%)	Top 10 clusters (%)
Characteristic loci	256 (27 after PCA)	78.71	99.01	100
Zoning	64(8 * 8)	52.3	58.43	65
Invariant moments	7	49.57	59	68.32
Fourier descriptors	30	79	83.41	91.33

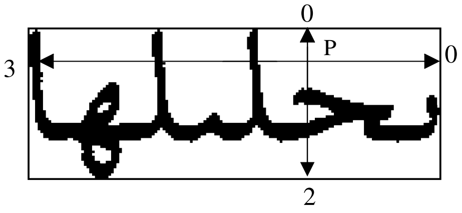


Fig. 5. Calculating characteristic loci features, $(2300)_4 = (176)_{10}$.

We employ the *k*-means algorithm with Euclidian distance to cluster the subwords into N_c clusters. The proper number of clusters is obtained by finding an optimum N_c based on entropy criterion. The mean of each cluster is used as its entry in the pictorial dictionary.

4.1. Characteristic loci

To describe the shape of subwords we tested different features: Fourier descriptors, invariant moments (Gonzalez and Woods, 2002), characteristic loci (Glucksman, 1967) and zoning and found out that the characteristic loci is the best among them.

Table 2 represents the results for clustering 5000 subwords of a test set with these features. As seen in this table, using zoning and moment features did not result in good partitions. Classification in cluster level with Fourier descriptors shows good results for the first cluster, but at most 91.53% classification rate can be reached for the first 10 clusters. To achieve 100% classification rate by Fourier descriptors, we should select the first 20 clusters. Thus we selected characteristic loci features to describe shape of subwords.

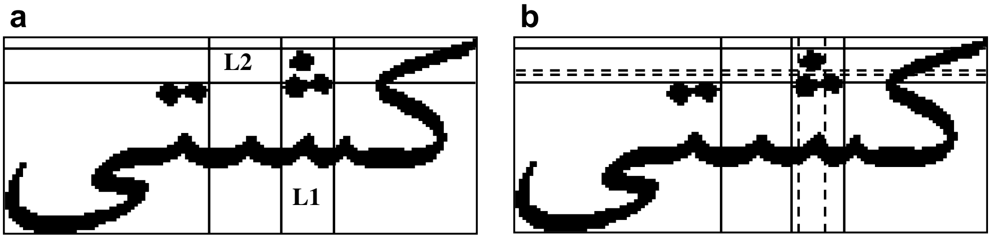


Fig. 6. (a) A sample subword with pen thickness of 6 pixels, loci "L1" and "L2" were formed considering smoothing in feature extraction. (b) Each of these two loci would have been divided to 3 smaller loci without smoothing.

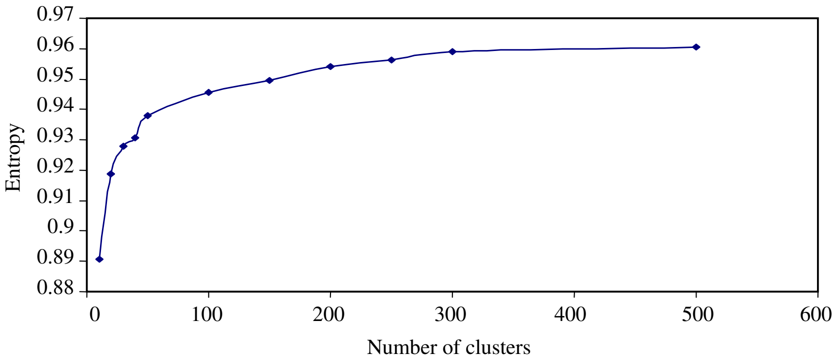


Fig. 7. Entropy measure vs. number of clusters.

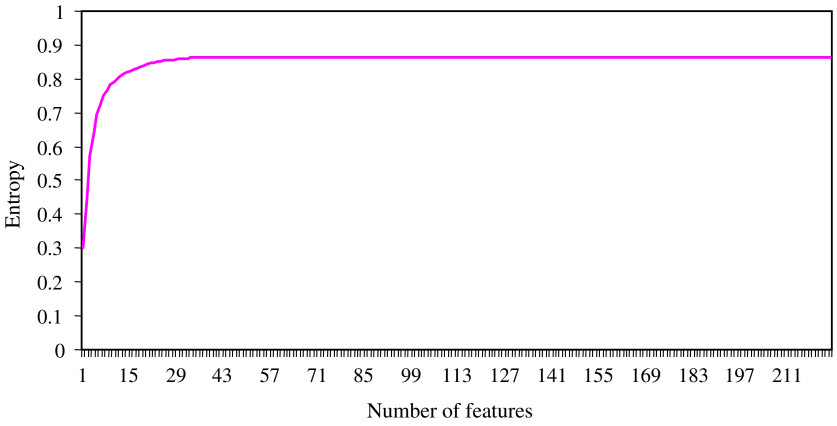


Fig. 8. Entropy for 300 clusters with different number of features.

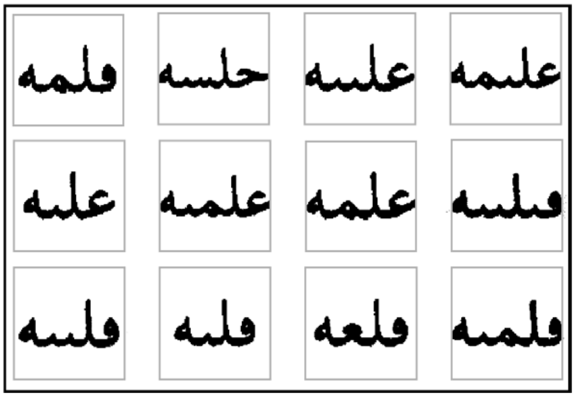


Fig. 9. Some members of a cluster of single-font subwords.

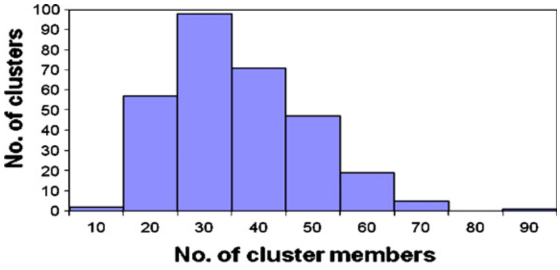


Fig. 10. Histogram of number of clusters as a function of number of their members for single font.

Characteristic loci features are usually defined in horizontal and vertical directions (Glucksman, 1967). In computing feature vectors, we assign a number to each background pixel as shown in Fig. 5. The features are computed according to the number of intersections with the subword body in right, upward, left and downward directions. In previous works, the characteristic loci method has been applied for digit and isolated letter recognition. In these applications, to reduce the feature dimension the maximum number of intersections has been limited to 2 (Glucksman, 1967; Knoll, 1969). The shapes of Farsi subwords could not be described perfectly by this limitation (Azmi et al., 2001). Therefore, we extended this limitation to 3. Then, for each background pixel, a four digit number of base 4 is obtained. For instance, the locus number of point P in Fig. 5, is $(2300)_4 = (176)_{10}$. The locus numbers are between 0 and 255. This is done for all background pixels. In this case, dimension of the feature vectors becomes 256. Each element of this vector represents the total number of background pixels that have locus number corresponding to that element. For example, 56th element of this vector represents the number of background pixels with locus number of 56. Features are normalized by dividing them by the total number of background pixels.

Table 3
Classification results in cluster level, for single font clustering

Number of closest clusters	1	5	7
Classification rate in cluster level (%)	80.69	97.52	100

To make the feature extraction more tolerable to noise, two modifications were made. Intersections with strokes smaller than half the pen thickness were ignored. Also, white spaces smaller than 0.75 of the pen thickness were ignored in feature extraction (see Fig. 6).

4.2. Finding the proper number of subword clusters

To build a pictorial dictionary of subwords, we use the k-means algorithm to cluster all subwords. The means of these clusters serve as dictionary entries.

The proper number of clusters was found by entropy measure in Eq. (1) (Pal and Dutta, 1986).

$$H = \frac{1}{N\log(N_c)} \sum_{j=1}^{N} \sum_{i=1}^{N_c} [D^{-1}(j,i)\log(D^{-1}(j,i))] \tag{1}$$

where, N is the total number of subwords, N_c is the number of clusters, and $D(j,i)$ is the distance of jth subword from the mean of ith cluster.

For a given N_c, if we have a crisp partition we have the highest information, i.e. the minimum entropy. Therefore,

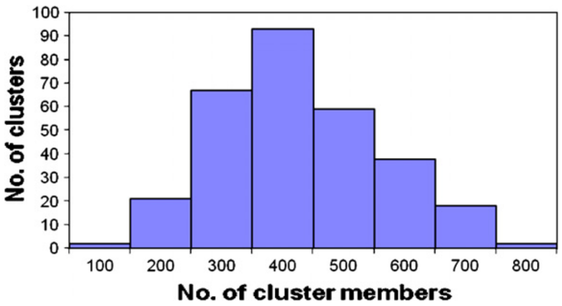


Fig. 11. Histogram of number of clusters for multiple fonts.

a partition with the minimum entropy is regarded as a good partition with compact clusters.

In our case, given a number of clusters, N_c, the best partition with minimum entropy was found and its entropy was saved. Changing N_c from 10 to 550, the entropy of best partitions for each N_c was calculated and plotted as a function of N_c (Fig. 7). For N_c larger than 300, this value did not change. It means that, increasing the number of clusters may not increase the information of clustering. Partitioning produced by clustering subwords using characteristic loci features contains a lot of information about the shape of subwords. This information increases as N_c increases. But increasing N_c may cause unwanted fractioning in clusters. Therefore, compact clusters may be divided to undesirable ones. This may not increase the information about subword clusters and the entropy of partition remains constant (Fig. 7).

Clustering of subwords was done for different number of clusters, using all 256 features, and the entropy criterion was computed. According to Fig. 7, clustering the subwords into more than 300 clusters did not change the entropy significantly. Therefore, we took 300 clusters for our work.

Table 4
Classification results in cluster level, for multifont clustering

Number of closest clusters	1	5	10
Classification rate in cluster level (%)	78.71	99.01	100

4.3. Feature reduction

33 out of 256 characteristic loci features were zero for all subwords of the training set. These features were omitted. PCA was used to reduce the dimension on the remaining 223 features. We used entropy criterion similar to Eq. (1), to choose the appropriate features. That is, using first feature, corresponding to the largest eigenvalues, the subwords were clustered to 300 clusters and the entropy of clusters was computed. Then using the first 2 features, corresponding to first 2 largest eigenvalues, the subwords were clustered to 300 clusters and the entropy of clusters was computed. This operation was repeated until clustering with all 223 features. The entropy of clusters vs. the number of features did not improve significantly for more than 27 features (Fig. 8). Therefore, we took these features for our work.

5. Using the pictorial dictionary in the recognition of subwords

In this section, some experimental results on the use of the proposed pictorial dictionary for the recognition of single and multifont subwords are reported and discussed.

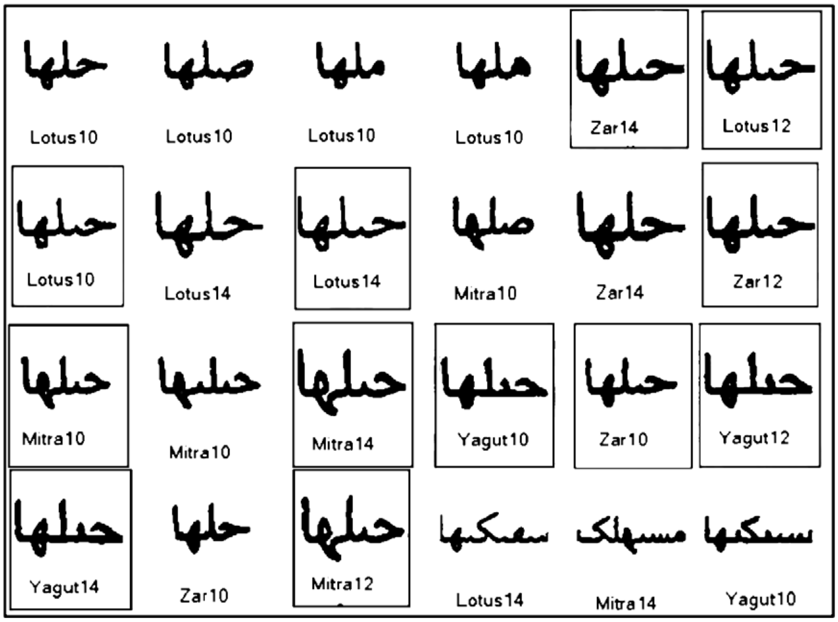
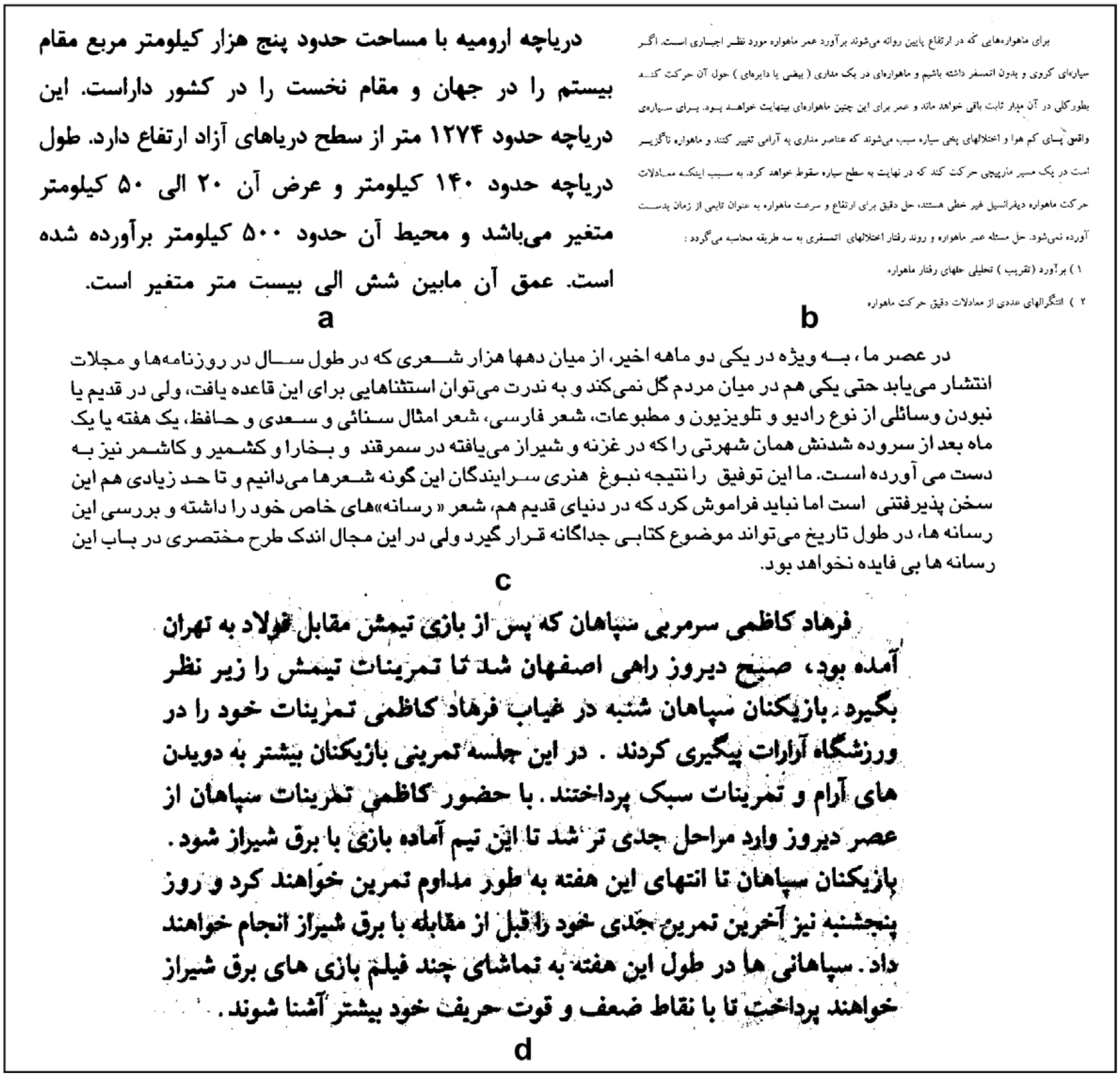


Fig. 12. Some members of a cluster of multifont subwords.



دریاچه ارومیه با مساحت حدود پنج هزار کیلومتر مربع مقام
بیستم را در جهان و مقام نخست را در کشور داراست. این
دریاچه حدود ۱۲۷۴ متر از سطح دریاهای آزاد ارتفاع دارد. طول
دریاچه حدود ۱۴۰ کیلومتر و عرض آن ۲۰ الی ۵۰ کیلومتر
متغیر می‌باشد و محیط آن حدود ۵۰۰ کیلومتر برآورده شده
است. عمق آن مابین شش الی بیست متر متغیر است.

برای ماهواره‌هایی که در ارتفاع پایین روانه می‌شوند برآورد عمر ماهواره مورد نظر اجباری است. اگر
سیاره‌ای کروی و بدون اتمسفر داشته باشیم و ماهواره‌ای در یک مداری (بیضی یا دایره‌ای) حول آن حرکت کند
بطورکلی در آن مدار ثابت باقی خواهد ماند و عمر برای این چنین ماهواره‌ای بینهایت خواهد بود. برای سیاره‌ی
واقعی بـای کم هوا و اختلالهای پخی سیاره سبب می‌شوند که عناصر مداری به آرامی تغییر کنند و ماهواره ناگزیر
است در یک مسیر مارپیچی حرکت کند که در نهایت به سطح سیاره سقوط خواهد کرد. به سبب اینکه معادلات
حرکت ماهواره دیفرانسیل غیر خطی هستند، حل دقیق برای ارتفاع و سرعت ماهواره به عنوان تابعی از زمان بدست
آورده نمی‌شود. حل مسئله عمر ماهواره و روند رفتار اختلالهای اتمسفری به سه طریقه محاسبه می‌گردد :

۱) برآورد (تقریب) تحلیلی حلهای رفتار ماهواره

۲) انتگرالهای عددی از معادلات دقیق حرکت ماهواره

در عصر ما، بـه ویژه در یکی دو ماهه اخیر، از میان دهها هزار شـعری که در طول سـال در روزنامه‌ها و مجلات
انتشار می‌یابد حتی یکی هم در میان مردم گل نمی‌کند و به ندرت می‌توان استثناهایی برای این قاعده یافت، ولی در قدیم یا
نبودن وسائلی از نوع رادیو و تلویزیون و مطبوعات، شعر فارسی، شعر امثال سنائی و سـعدی و حـافظ، یک هفته یا یک
ماه بعد از سروده شدنش همان شهرتی را که در غزنه و شیراز می‌یافته در سمرقند و بـخارا و کشمیر و کاشـمر نیز بـه
دست می آورده است. ما این توفیق را نتیجه نبـوغ هنری سـرایندگان این گونه شـعرها می‌دانیم و تا حـد زیادی هم این
سخن پذیرفتنی است اما نباید فراموش کرد که در دنیای قدیم هم، شعر « رسانه»های خاص خود را داشته و بررسی این
رسانه ها، در طول تاریخ می‌تواند موضوع کتابـی جداگانه قـرار گیرد ولی در این مـجال اندک طرح مـختصری در بـاب این
رسانه ها بی فایده نخواهد بود.

فرهاد کاظمی سرمربی سپاهان که پس از بازی تیمش مقابل فولاد به تهران
آمده بود، صبح دیروز راهی اصفهان شد تا تمرینات تیمش را زیر نظر
بگیرد. بازیکنان سپاهان شنبه در غیاب فرهاد کاظمی تمرینات خود را در
ورزشگاه آزارات پیگیری کردند . در این جلسه تمرینی بازیکنان بیشتر به دویدن
های آرام و تمرینات سبک پرداختند. با حضور کاظمی تمرینات سپاهان از
عصر دیروز وارد مراحل جدی تر شد تا این تیم آماده بازی با برق شیراز شود.
بازیکنان سپاهان تا انتهای این هفته به طور مداوم تمرین خواهند کرد و روز
پنجشنبه نیز آخرین تمرین جدی خود را قبل از مقابله با برق شیراز انجام خواهند
داد. سپاهانی ها در طول این هفته به تماشای چند فیلم بازی های برق شیراز
خواهند پرداخت تا با نقاط ضعف و قوت حریف خود بیشتر آشنا شوند.

Fig. 13. Some test images from a book (a), a conference proceedings (b), a journal (c) and a newspaper (d).

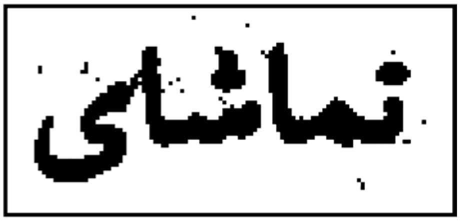


Fig. 14. An example of touching subwords.

5.1. Single-font dictionary

The most popular Farsi font, Lotus of size 14, was used for single-font experiment. 9445 subword bodies were clustered to 300 clusters by the *k*-means algorithm. The minimum and maximum numbers of subwords in these clusters were 6 and 75, respectively. In Fig. 9 some subwords of a sample cluster are shown. The subwords with similar shapes are grouped together. All the subwords of this cluster have a similar ascender, beginning letter and the same ending letter.

The histogram of number of clusters as a function of number of their members for single font clustering is shown in Fig. 10. Ninety-seven clusters have between 20 and 30 members. About 92% of clusters have less than 50 subwords and just 25 clusters have more than 50 members.

Experiments showed that the minimum mean-distance classifier finds the correct subword within at most 7 closest clusters. Therefore, the number of subwords candidates that one should compare to input subword are reduced from 9445 to at most $75 * 7 = 525$ subwords.

In a test, a set of 1000 new subword samples of font Lotus, size 14, were classified in cluster level. Results are shown in Table 3. In this table, the recognition rate indicates that, for example, if we take 5 closest clusters to the input subword, the probability of having the correct subword within these clusters is 97.52%. Choosing 7 closest clusters to an unknown subword guarantees the existence of the correct answer among the subwords of those clusters.

5.2. Multifont dictionary

Multifont clustering for 4 fonts in 3 sizes, resulted in clusters with minimum and maximum number of 74 and 800 subwords, respectively. The histogram of number of clusters as a function of number of their members for multifont clustering is shown in Fig. 11. Most frequent numbers of cluster members were between 300 and 400. Similar to single font clustering, compact clusters with very similar members were created.

Experiments showed that the minimum mean-distance classifier finds the correct subword within at most 10 closest clusters. Therefore, subwords candidates that one should compare to input subword are reduced from 113,340 to at most $800 * 10 = 8000$ subwords. Some sample subwords of a cluster are shown in Fig. 12. As we can see, all samples of the subword "حيـلـــهـا" in fonts Zar, Lotus, Mitra and Yagut are located in this cluster.

In a test, a set of 5000 new subword samples were classified in cluster level. Results are shown in Table 4. Considering 10 closest clusters to the input subword, the minimum, maximum and mean numbers of candidates were 2368, 5873 and 4060, respectively. Therefore the search domain for an unknown subword was reduced to 4060 subwords.

5.3. An experiment on noisy images

We tested our algorithm on a set of images with different qualities. Images from different resources like newspapers, books, conference proceedings and journals were selected. The image set consists of 15,000 subwords. Some sample images are shown in Fig. 13.

15,000 subwords with different qualities and fonts are classified to 10 closest clusters using Euclidian distance. The correct classification rate was 99.8%. Touching subwords were the main source of the errors. An example of these errors is shown in Fig. 14. In this figure, subwords "شـا" and "ى" are connected. Thus, a shape is produced that is not in our pictorial dictionary. This shape is assigned to an incorrect cluster.

6. Conclusion

In this paper, to build a pictorial dictionary of printed Farsi subwords, we used the characteristic loci features, compressed by PCA, to cluster these subwords. We employed the *k*-means algorithm with Euclidian distance. Initial points were selected uniformly from input data. The proper number of clusters was obtained based on entropy criterion. The mean of each cluster was used as its entry in the pictorial dictionary.

To evaluate the clustering results, a minimum mean-distance classifier was used to test a set of 5000 subwords, in 4 fonts and 3 sizes. 78.71, 99.01 and 100 percent of these subwords were in the first, first five and first ten closest clusters, respectively.

The proposed pictorial dictionary can be employed to reduce the search domain for subword recognition. This pictorial dictionary can also be used in a content-based document retrieval system.

References

Allam, M., 1995. Segmentation versus segmentation-free for recognizing Arabic text. Proc. SPIE 2, 228–235.

Amin, A., 1930. Handwritten Arabic character recognition by the I.R.A.C. system. In: Fifth Internat. Conf. Pattern Recognition, Florida, December, pp. 729–731.

Azmi, R., 1999. Recognition of omnifont printed Farsi text. PhD Thesis, Tarbiat Modarres University, Tehran, Iran (in Farsi).

Azmi, R., Kabir, E., 2001. A new segmentation technique for omnifont Farsi text. Pattern Recognition Lett. 22, 97–104.

Azmi, R., Kabir, E., Badie, K., 2001. An algorithm for clustering and recognition of omnifont Farsi subwords. Int. J. Eng. Sci., IUST 12 (1), 39–49 (in Farsi).

Badie, K., Shimura, M., 1980. Machine recognition of Arabic cursive script. Pattern Recognition Practice, 315–323.

Bazzi, I., Schwartz, R., Makhoul, J., 1999. An omnifont open-vocabulary OCR system for English and Arabic. IEEE Trans. Pattern Anal. Machine Intell. 21 (6), 495–504, June.

Ben Amara, N. Belaid, A., 1996. Printed PAW recognition based on planar hidden Markov models. In: 13th Internat. Conf. Pattern Recognition, vol. 2, Vienna, pp. 220–224.

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

Ebrahimi, A., Kabir, E., 2005. A two step method for the recognition of printed subwords. Iranian J. Electric. Comput. Eng. 2 (2), 57–62 (in Farsi).

Glucksman, H.A., 1967. Classification of mixed-font alphabets by characteristic loci. In: Proc. IEEE Comput. Conf., September, pp. 138–141.

Gonzalez, R.C., Woods, R.E., 2002. Digital Image Processing, Second ed. Prentice Hall.

Ho, T.K., Hull, J.J., Srihari, S.N., 1992. A word shape analysis approach to lexicon based word recognition. Pattern Recognition Lett. 13 (11), 821–826.

Huang, W., Tan, C.L., Sung, S.Y., Xu, Y., 2001. Word shape recognition for image-based document retrieval. In: Internat. Conf. on Image Processing, ICIP2001, Greece, pp. 1114–1117.

Hull, J.J., 1985. Word shape analysis in a knowledge-based system for reading text. In: The Second IEEE Conf. on Artificial Intelligence Applications, pp. 114–119.

Khorsheed, M.S., Clocksin, W.F., 2000. Spectral features for Arabic word recognition. Proc. ICASSP00 6, 3574–3577.

Knoll, A.L., 1969. Expriments with 'characteristic loci' for recognition of handprinted characters. IEEE Trans. Comput. C-18 (April), 366–372.

Mori, S., Suen, C.Y., Yamamoto, K., 1992. Historical review of OCR research and development. Proc. IEEE 80 (7), 1029–1058.

Pal, S.K., Dutta, D.K., 1986. Fuzzy Mathematical Approach to Pattern Recognition. Wiley Eastern Limited.

Parhami, B., Taraghi, M., 1981. Automatic recognition of printed Farsi text. Pattern Recognition 14, 395–403.

Powalka, R.K., Sherkat, N., Withrow, R.J., 1997. Word shape analysis for a hybrid recognition system. Pattern Recognition 30 (3), 421–445.

Razavi, S.M., Kabir, E. 2004. A database for on-line handwritten Farsi subwords. In: Sixth Conf. on Intelligent Systems, Kerman, Iran, pp. 218–225 (in Farsi).

Shahreza, M.S., Faez, K., Khotanzad, A. 1994. Recognition of handwritten Farsi numerals by Zernike moments features and a set of class-specific neural networks. In: Proc. ICSPAT-94, T0065as, USA, October, pp. 998–1003, 18-21.

Trier, O., Jain, A.K., Taxt, T., 1996. Feature extraction methods for character recognition – a survey. Pattern Recognition 29 (4), 641–662.

Zheng, L., Hassin, A.H., Tang, X., 2004. A new algorithm for machine printed Arabic character segmentation. Pattern Recognition Lett. 25, 1723–1729.