Journal of Engineering and Applied Sciences 12 (8): 2009-2017, 2017
ISSN: 1816-949X

Recognition of the Persian Typed Sub-Words with a Hierarchical Manner

Esmail Miri, Seyyed Mohammad Razavi and Naser Mehrshad
Department of Electrical and Computer Engineering, The University of Birjand, Birjand, Iran

Abstract: In this study, a simple way to recognize the Persian typed sub-words are provided. First, the search space is limited by using a few simple characteristics and according to the position of points and signs of input sub-words. Then by using the ofloci features, two sub-words with the smallest distance than input sub-word are selected. Also, two sub-words are selected by using zoning features and minimum distance criteria. To select final sub-word among following four sub-words, the number of text vertical intersection with field feature is used. This method was tested for Lotus font and 98.18% recognition rate is achieved for test data.

Key words: Search space reduction, Persian sub-words recognition, points and symbols, printed Persian sub-words, text recognition, vertical intersection characteristics

INTRODUCTION

In words recognition field three approaches of words isolation, the overall shape and combination of the two is used. Most optical recognition techniques of grapheme by breaking the word into its constituent letters and recognition of these letters recognize the word. This approach faced the problems because of the letters isolation problems and their accurate recognition in texts with a low image quality. Therefore, the approach base on none separating recognition it would be useful in such cases. On the other hand, a lot of psychology research is done about human reading which one of the results is that the human eye when reading a text line, do not move continuously from left to right but discretely from one focus to another focus point jumps. Research has shown that the time required recognizing a four-letter word is equal to the time it takes to recognize a separate letter. Studies have emphasized the importance of the overall form of the word in the process of recognition and with respect to the subject for the use of visual features in the word's surface, methods suggested (Adamek *et al*., 2007; Amin, 2000).

In the methods based on the overall shapes, generally, for recognition of sub-word after extracting the sub-word characteristics, features formed in the picture glossary search in the training phase. Therefore, by considering that we are faced with massive amounts of classes, offering ways to limit the scope of the search of the main challenges in the overall shape is based on overall shapes. Thus the information of overall shape usually is used to reduce the scope search in a hierarchical system (Ebrahimi, 2007; Khosravi and Ehsanollah, 2011; Madhvanath *et al*., 1999).

In addition of recognition systems, using word overall shape information in word recovery among word limited sets, dramatically reduces processing volume. Also, describe of overall shape is an efficient way to display the query words in the document images (Li *et al*., 2007; Bai *et al*., 2009; Lu and Tan, 2004; Rodriguez-Serrano and Perronnin, 2009; Rath and Manmatha, 2007).

Reduce the search space, in addition to reducing the amount of calculations required in the later stages will increase the final system accuracy of recognition. The aim of this study is to provide a simple way and yet effective method for recognition of Persian typed sub-words with lotus font in several steps. In the first step by determining the ratio of width to height of the sub-word (with signs and without signs), scope of search is limited to sub-words with a range of these ratios. In the second phase the ratio of the number of image black dots to image surface, the ratio of the number of image top half black dots to image black dots and the ratio of the number of image right half black dots to image black dots are calculated and the search scope is limited to sub-words with range of these ratios. In third step after passing through the first and second inhibitors with the respect to signals position, the scope of search is limited to sub-words their sigs position are identical with the input sub-word. Then, using each loci and zoning features and criteria of the smallest distance of two closer options to input sub-word selected and in the final stage, the sub-word chosen from the four optionsselected by using features of the number of text vertical intersection with field. Figure 1 shows the major components of the proposed process. The database used in this study sets the following 12700 common words in Persian with which lotus 14 font size written and printed and scannedwith a resolution of 400 dpi (Ebrahimi and Kabir, 2008). To create test samples, 1000 sub-words are randomly selected from database sub-words. This sub-words have been printed

Corresponding Author: Esmail Miri, Department of Electrical and Computer Engineering, The University of Birjand, Birjand, Iran

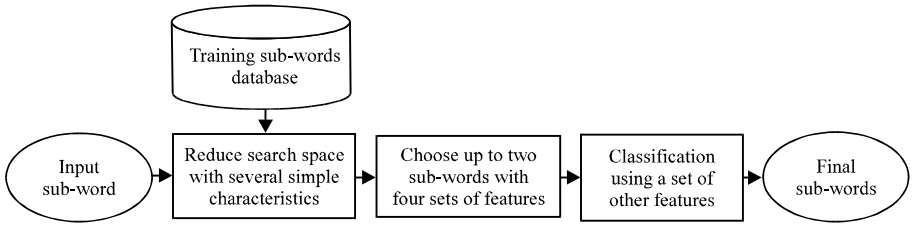

Fig. 1: Main components of proposed process

with lotus font and in three font sizes (14, 12 and 10). The 12 and 14 font sizes sub-words have been scanned in 300 and 200 dpi resolution degree and 10 font sizes sub-words have been scanned in 300 dpi resolution degree. In total, 5000 samples have created for test.

Literature review: Continuity letters in some lines and some styles of writing, makes the segmentation of words more complicated and will causes the researchers are more likely to have overall shape-based methods. Farsi, Arabic writings as well as English hand writings include this category. Several studies have been conducted on the recognition of continuous lines (Rehman and Saba, 2012).

By Madhvanath and Govindaraju (2001) various methods are provided to hand written word recognition-based approach and Arabic words recognition techniques with both segmentation-based and overall shape-based approach are reviewed by Lorigo and Govindaraju (2006).

In overall shape-based methods, generally after extracting sub-word features, these features have been searchedin picture dictionary formed in training step. Due to large number of classes, select the proper characteristics that may provide a breakdown of the number of classes is one of the bottlenecks of this method. Various studies have used different methods for extracting the characteristics. Feature extraction nationwide of words form using a variety of structural and statistical descriptors like horizontal and vertical affections, number and position of dots, sigs, the risings, scrollable, holes, bezels, high and low profile form pixel density and nationwide. Converting have been studied in the literature (Madhvanath and Govindaraju, 2001).

In previous studies about Farsi language in different ways efforts have been made to reduce the scope of search. Initially, Ebrahimi to reduce scope of search divided sub-words of database into 300 clusters with different fonts and size and features of loci. In classification step, the input sub-word is compared with central of mentioned clusters and the closest clusters by using of Fourier descriptors features are searching quickly.

Davoudiand have been used the clustering with locifeatures and increasing the range of certainty on selected cluster base on features of local shape toward reducing the number of selected clusters and as a result to more reduce of search scope.

In another research, Fathi (2011) benefit from index letters to reduce the search scope. In this research, the first and the last special letters are recognized without segmentation and at the classification stage; the input sub-word is explored just among these special letters.

The code of points is also another technique done (Alibeigi, 2017). In this study, the input sub-word is searched through the sub-words with similar codes. In this study in classification step the input sub-word, searches just among sub-words which have identical dots, thus search scope reduces dramatically.

In another research clustering, the signal position code and the ratio of width to height of sub-words have been used for restricting search space. In this research in the first step by extracting the simple features of horizontal and vertical profiles, search space has been restricted into number of selected clusters. In the second step by determining the ratio of width to height of sub-word, search scope is restricted into sub-words with the scope of this ratio. In the third step with the respect to signs position only the sub-words which have the identical signal position with input sub-word are searched. With the proposed method the search space is reduced to an acceptable level.

MATERIALS AND METHODS

The general structure of the proposed method: In the first stage, the required geometric features of database sub-words include ratio of width to height of sub-word (with signs and without signs), the ratio of the number of image black dots to image surface, the ratio of top half black dots of image to black dots of image and the ratio of the number of right half black dots of image toblack dots of image are extracted. Then, the sign code position and loci features and zoning features of database sub-words are extracted. The test can be done to reduce the search space in a multi-level way (Fig. 2).

At the first level ratio of width to height (with signs and without signs) of input sub-word is compared with

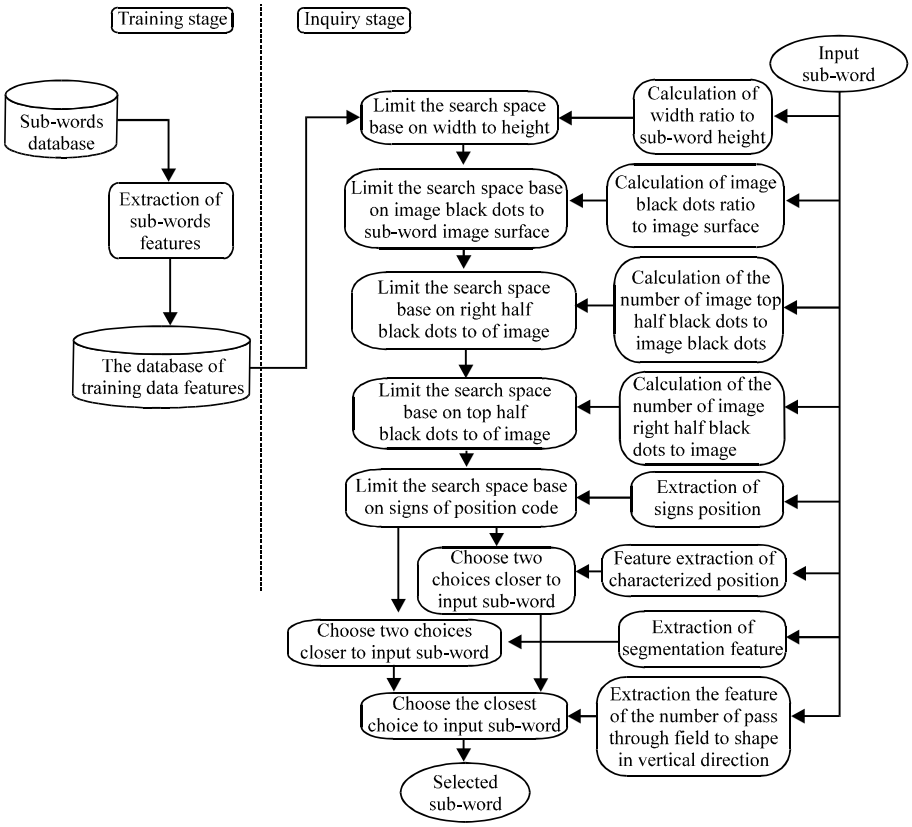

Fig. 2: Proposed method structure

ratio of width to height of database sub-words and the sub-wordslocated inspecified restriction of this ratio are selected.

In next levels the same stages with the ratio of the number of black dots of image to surface of sub-word image, the ratio of the number of top half black dots of image to black dots of sub-words and the ratio of the number of right half black dots of image to black dots of sub-word image repeats in the previous restricted space and finally to reduce the search space the signals position is used.

Firstly, in recognition stage, by using loci features and criteria of the smallest distance a pair of options closer to input sub-word is selected. This step is repeated with zoning features. In the final stage, recognition of selected sub-word out of four selected sub-word of previous using features of the number of text vertical intersection with field toward vertical and neighborhood classifier is done.

One major difference with previous methods, the method used is simple and multi-stage process. In this study, instead of trying to sophisticated feature selection or using techniques such as clustering, in search space reduce stage using simple inhibitors like the ratio of width to height of sub-word, the ratio of the number of image black dots to sub-word image surface, the ratio of the number of top half black dots of image to black dots of sub-word image, the ratio of the number of right half black dots to sub-word black dots of image and the signals position, significant results in the search space reduction achieved. And finally, to classification, two classified stages are used. In following, each of the above steps is described.

Ratio of width to height of the word: Simple and useful restriction that has been used to reduce the search space is using the ratio of width to height of the input sub-word. The main character of this inhibitors is due to its relative structure is independent of the font size. In this study, to reduce the search space and at the same time cover the entire sub-words performed numerous tests and finally the search space of sub-words are restricted to the sub-words which have the ratio of width to height between 0.7-1.3 rather than ratio of width to height of input sub-word (with signs and without signs). It shows how to calculate the ratio of width to height with signs and without signs schematically, the ratio is 1.957 in without signs mode and equals 1.343 in signs mode. The ratio of width to height of a sample sub-word is seen. With a first look at the Table 1 and given the obvious

Table 1: The ratio width to height with signs and without signs of some sub-words

Sub-word: the ratio of width to height	
With signs	Without signs
1.15	2.25
1.46	2.48
1.25	1.68
2.25	2.88
2.53	3.38
2.12	2.87
3.45	3.67
2.02	2.70
2.73	3.02

differences between sub-word in terms of the usefulness of this feature, the use of this character will appear in the search space reduction, e.g., for lotus font the ratio width to height of sub-word database is in intervals from 0.136-9.36 without signs and varies 0.136-5.3 with signs.

The ratio of the number of points: Another simple feature that has been used in this study to reduce the search space is using the spatial distribution of black dots in the main body of sub-word by calculating, the ratio of the number of image black dots to sub-word image surface, the ratio of the number of top half black dots of image to black dots of sub-word image, the ratio of the number of right half black dots to sub-word black dots of image.

In previous research including, Ebrahimi (2007), Fathi (2011) and Alibeigi (2017) code of sub-word sings are used in different stages which causes high errors in writingwith small font size and lowseparation accuracy, thus, in this study to solve this problem just signs position code is used to limit the search space. Some properties of signs that can be deduced from this table include:

- With regard to the specific form of three points combination of them in the first step seems logical
- In two neighbors signed letters vertical position of the sign center in the most fonts are different
- The distance between two points on one letter is less than two points on twoneighbors letters
- Sometimes identifying baseline to determine high or low sign is a hard work especially when investigating of separated sub-words is different and with error. For example, it is possible in the sub-word "ﮐ" baseline under the "ﮔ" to be determined
- Signs distance is a function of the font size

The properties listed and the method of trial and error the following rules were designed to extract the signs position code. The stages of sign code extraction are:

- The labeling of components and separate body of sub-word (the largest component)
- Extract the pen width (the most frequent thickness in the body of the sub-word)
- Incorporate points of a letter: in the beginning. The sub-word signs are arranged from left to right then the first sign with its left sign (if any) is checked in conditions
- If the horizontal distance of two sign is <1.7 times the width of the pen
- If the vertical distance of frames of two signs is <1.7 times the width of the pen

If the above conditions are met simultaneously two signs are combined together and consider as one and above stages repeat again with combination component. The same stages are performed for other signs, respectively. The components related to 3-pointed and 2-pointed are combined after doing this process.

Signs positioning: To determine sign up or down if consider the coordinates ofupper left corner of the frame $X_{b\ min}$, $Y_{b\ min}$ and the coordinates of lower right corner of the frame $X_{b\ max}$, $Y_{b\ max}$ and use $X_{c\ max}$, $Y_{c\ max}$, $X_{c\ max}$, $Y_{c\ max}$ for the signs related to coordinates of lower right corner, the frame center coordinates will be like this:

$$X_{c\ mean} = \left(\frac{X_{c\ min} + X_{c\ max}}{2} \right)$$

$$Y_{c\ mean} = \left(\frac{Y_{c\ min} + Y_{c\ max}}{2} \right)$$

If, $X_{b\ min} \leq X_{c\ mean} \leq X_{b\ max}$ move upward from the center of the of ($X_{b\ max}$, $Y_{c\ mean}$) if, faced the body the sign position frame if faced the sub-word body consider the signs position down, otherwise, the sign position is considered up. If $X_{c\ mean} > X_{b\ max}$ move upward from a point with coordinates is down and otherwise, sign position is considered up. If $X_{c\ mean} < X_{b\ min}$ move upward from the point with coordinates of ($X_{b\ min}$, $Y_{c\ mean}$), if meet the body the sign position is down and otherwise the sign position is considered up.

In next stage a sign position code is allocated to each sub-word. It should be noted that there are many errors in diagnosis of signs but the algorithm used in determining the position of the sign with only rare errors may occur due to the poor quality of scanned images.

Assign code: Code assign method to sub-words is this that the considerable word is investigated right-to-left. In each letter if one of the signs (one pointed, two-pointed (combined), three pointed (combined), Sarkesh " ", Mad "ٓ", Tashdid "ّ", Hamza "ٔ") is located top of the body, code 1 assigns that letter and if each of the signs is located under the body takes code 2. Otherwise no code is assigned to that letter and finally by juxtaposing letters code left-to-right, the sub-word code achieved. The

Table 2: Position sign code of some sub-words

Sub-words	
Sign position codes (1)	Sign position codes (2)
221	2111
211	2111
21	21121
212	211
2112	2111

Table 3: Frequency of position sign code of database sub-words

Codes	Number	Codes	Number	Codes	Number	Codes	Number	Codes	Number	Codes	Number
0	856	2122	17	12112	4	22111	8	112121	1	221111	1
1	2139	2211	86	12121	4	22112	11	112221	1	221211	1
2	1387	2212	36	12211	10	22121	13	121111	1	221212	1
11	1831	2221	28	12212	2	22122	1	121121	1	221222	1
12	1166	2222	9	12221	4	22211	3	121221	1	222121	1
21	1416	11111	16	12222	1	22212	4	122221	2	1121111	1
22	712	11112	8	21111	4	111112	1	211111	2	112211	1
111	627	11121	26	21112	2	111121	1	211121	1	2112111	1
112	379	11122	2	21121	13	111122	1	211211	3	-	-
121	667	11211	18	21122	3	111211	3	211212	3	-	-
122	212	11212	4	21211	11	111212	1	211221	1	-	-
211	482	11221	19	21211	2	111221	2	212112	1	-	-
212	230	12111	16	21221	4	112111	8	212121	2	-	-

Table 4: The results of search space reduction

	Percent of search space reduction		
Types of restrictor	2500 initial data (%)	2500 test data (%)	5000 data (%)
The ratio of height to width of sub-word	62.00	62.00	62.00
The ratio of image black dots to image surface	45.20	44.20	44.70
The ratio of the number of top half black dots to black dots	60.00	60.60	60.30
The ratio of the number of right half black dots to black dots	20.00	20.60	20.30
Signs position	90.70	90.70	90.70
The ratio of height to width+the ratio of the number of dots	93.80	93.40	93.60
All	99.40	99.32	99.36

sub-word that is no sign at the bottom and above code will be assigned zero. Again is recalling numbers and proportions used in the extraction process of sign position are achieved with try and error and attempt to provide a general design and formula. Table 2 shows some of their sub-words and position code of signs extraction. The frequency of different codes in database is shown in Table 3. As seen in Table 3, the most frequency achieved in a set of 86 codes by 2139 cases (16.8%) is related to code 1 (up sign). 46 out of 86 codes have the frequency <10 while 22 codes of it contain only one member. The analysis of information provided by this Table 4, the high performance of using these inhibitors is clearly visible. As a corrective point with respect to the connection of Sarkesh "ـ" to body is possible in some of fonts or low separation degrees. To avoid the error for letter "ک" both codes are considered without code and an up code which this adds the number of database elements from 12700-13882.

Extraction of zoning features: Shows how to extract zoning features for a sub-word in this study. in order to zoning features extraction in each direction of shape divided as equally as may be into eight parts (roundthe divisions results), In each block number ones (houses occupied by the word) is summed and normalization divides on the total number ofblocks home. By this way the 64 features per each sub-word has been achieved.

Features of loci: One of the used features in recognition stage in this study is loci features. Loci features usually defined in terms of vertical and horizontal or diagonal directions. Loci features vectors are calculated like this that a number is assigned to each point of image field. This number, according to the vertical and horizontal lines drawn from that point in four-direction top, down, right and left, cuts the sub-word in several points, is calculated. In this study with respect to complicated Fig. 3 of Farsi sub-words the body cut number is limited to 3; therefore, a four-digit number in base 4 can be obtained. These numbers are used to display loci features of equivalent base 10. Loci features vectors have 256 members in this status which each one shows its frequency number in the field of image, in other word the relative characterized position surface. To normalize these features, vectorelements divideon the number of white points of image field. The general form of the sub-word with the frequency vector of characterized position features is displayed in the picture.

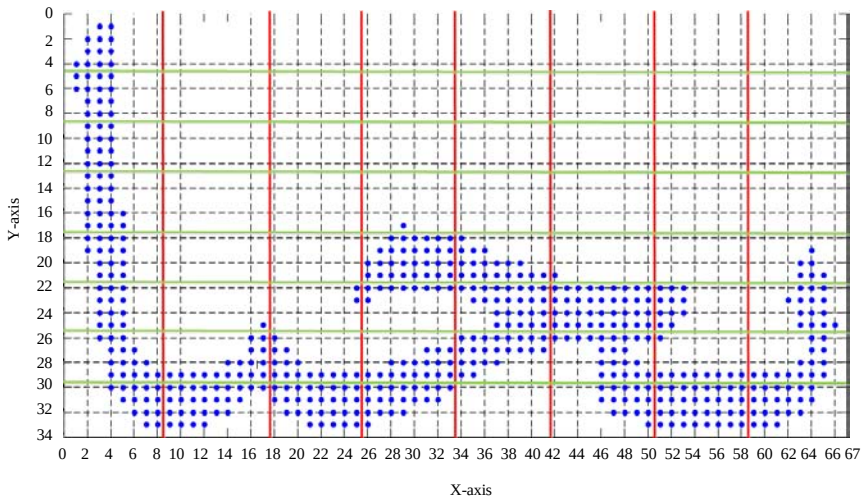

Fig. 3: Extraction of segmentation features

Extraction of features of the number of text vertical intersection with field: In this study, the calculation of the number of text intersection with field in vertical direction are the final features used in the recognition stage in order to sub-word detection at last from 4 proposed sub-words of final stage. According to all of the following words are not the same width, obtained feature vector is normalized to the number of elements.

In Fig. 3, the text intersection place with field is specified with white points which achieved bycounting points in each column of features. If the image width is N pixel and we have M features we do this, we produce M real number at first which the first one equals 1 and the last of them equals N. we divide between 1 to N into M-1 equal intervals and gain M-2 other real word and round these numbers until have the elements number that should choose from N initial element.

If N is greater than or equal to M the numbers of elements would not be repetitive but if M is greater than or equal N we also have repetitive numbers that select thesamplesfor several times.

For example if N = 70, M = 10. If between 1 and 70 are divided into nine equal parts numbers we have: 1, 8.67, 16.33, 24, 31.67, 39.33, 47, 54.67, 62.33 and 70. If these numbers are rounded the following numbers achieve: 1, 9, 16, 24, 32, 39, 47, 55, 62 and 70. Which shows the number of vectors is features that should be selected. But for example if M = 10 and N = 7, we have these numbers: 1, 1.67, 2.33, 3, 3.67, 4.33, 5, 5.67, 6.33 and 7. And if round above numbers the following numbers achieve: 1, 2, 2, 3, 4, 4, 5, 6, 6 and 7. As can be seen the numbers 2, 4, 6 repeated. So, the elements of the 2nd, 4th and 6th two times each and the rest will be selected only once.

Classifier: After the search space reduction which leads to reducing search space from 12700-82 sub-words (on average) its turn to use classifier in order to determine the final sub-word. Choose a feature extraction method is proposed as an important factor in recognition system performance. To identify letters and digits, zoning features, loci features, geometric moments, zernike moments, quick descriptors, constant torque, histogram view, describing graphs, etc. have been used. Optional features must be such that in addition to expressing sub-word characteristics may also provide separation between different sub-words. In this regard, after studying and testing different features in this study, three features loci, zoning and the number of vertical intersection text field is used. With regard to complication of Farsi sub-words none of above features in the first option deal to sub-word recognition with high rate of limited sub-words set. Thus recognition considered in two stages. In the first stage number of the closest sub-words to input sub-word are selected then in the second stage the final sub-word is selected among selected sub-words of before stage with a classifier.

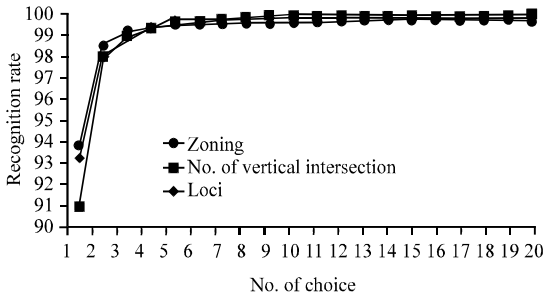

Fig. 4: The correct recognition by increasing the number of words with different characteristics

As can be seen in Fig. 4 by selecting the nearest sub-word to input sub-word, the number of true

options for loci feature, zoning and the number of vertical intersection text field are: 93.3, 93.9 and 91%, respectively. To achieve acceptable true choice the first 12 choices of loci feature or the first 15 choices with number of vertical intersection text field and over 20 closer choices of zoning feature are selected. So to avoid of the number of selected sub-words increasing which causes more complicated final stage by multiple combination experiments of zoning and loci features as general features for first stage and features of the number of vertical intersection text field which has more partial view selected for second stage. By selection of two first choice of each one of both zoning feature and loci a satisfy result have been obtained. In this stage of classifier the closest neighborhood with K = 1 and Euclidean distance criterion (best choice according to experiments) is used and the desired result is achieved. The result of first stage of classification limits to choice scope up to 4 sub-wordswhich are very similar.

In the last stage of recognition to detect final sub-word from the number of vertical intersection text field features which have more partial view are used. shows the recognition hierarchy of the sub-word used in the confined space of limiters, four selected sub-word with two loci feature and zoning and final recognized sub-word, respectively.

RESULTS AND DISCUSSION

Experiments and analysis of the results: The database used in this study sets 12700 common sub-words in Persian which is printed by lotus font and written by 14 font size and have been scanned with a resolution of 400 dpi and to create an experimental sample, 1000 following word database randomly selected among the following words. These sub-words are printed by lotus font and three font size of 10, 12 and 14 and have been scanned by 200, 300 dpi resolution degree and include 5000 samples. The 2500 samples out of 5000 are applied to set the parameters and other 2500 are test samples. As explained, reducing search space happens in several stages in this study. In the first level, the ratio of width to height of extracted sub-word and database sub-words is compared and the sub-words are selected which are located in the specific area of the ratio. In the next levels the same procedure for the ratio of the number of image black dots to the image surface of sub-word, the ratio of the number of top half black dots of image to black dots of sub-word image and the ratio of the number of image right half black dots to image black dots of the sub-word repeated in previous limited space and finally to reduce the search space, signs position have been used. Then in recognition stage first four nearest sub-word to the input sub-word are selected (by aid of loci features and zoning featuresand the closest

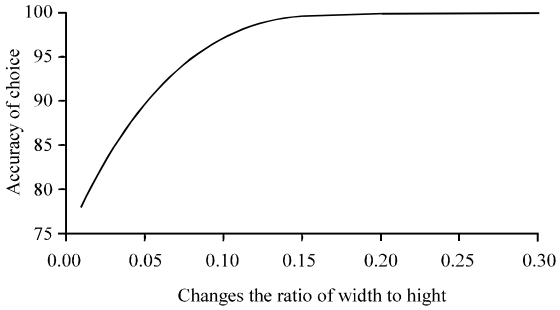

Fig. 5: Selected sub-word accuracy changes by increasingratio of width with height (test data)

neighborhood). In the final stage of recognition of selected sub-word out of four selected sub-word of previous stage happens using the number of intersection text field features in vertical direction and the closest neighborhood.

The first stage of the search space reduction is using ratio of width to height as search space limiter. Shows the choice accuracy changes diagram to change of this ratio. There are the first selected intervals in this Fig. 5 and 6 with ratio of width to height between 0.84-1.26 equals ratio of width to height of input sub-word in selected clusters set which 0.01 is added to this intervals in each stage. According to this chart by selecting the ratio of width to height in the range of 0.7-1.3, choice accuracy is up to 100%. By using this limitation, search space reduces from 12700-4826 sub-words; therefore by maintaining 100% accuracy62% reduction appears on search space.

A similar process with ratio of width to height passed for three simple features of number of image black dots to image surface, the ratio of image top half black dots to image sub-word black dots and the ratio of the number ofimage right half black dots to image sub-word black dots which according to experiments result with 100% accuracy and in other hand maximum limitation with the number of image black dots limiter to image surface, sub-words should be selected that the ratio is in the range of 0.75-1.2 for them.

This area is between 0.5 and 1.9 for ratio of the number of image top half black dots features to image black dots of sub-word (for 100% accuracy in first 2500 database) which causes a reduction about 40%. For ratio of the number of image right half black dots to image black dots of sub-word the area is between 0.6-1.5 (for 100% accuracy in 2500 first data) and causes 20% reduction of search space.

In next stage reduction of search space tested by using signs position which this stage by using designed algorithm with 100% accuracy in 2500 first data causes

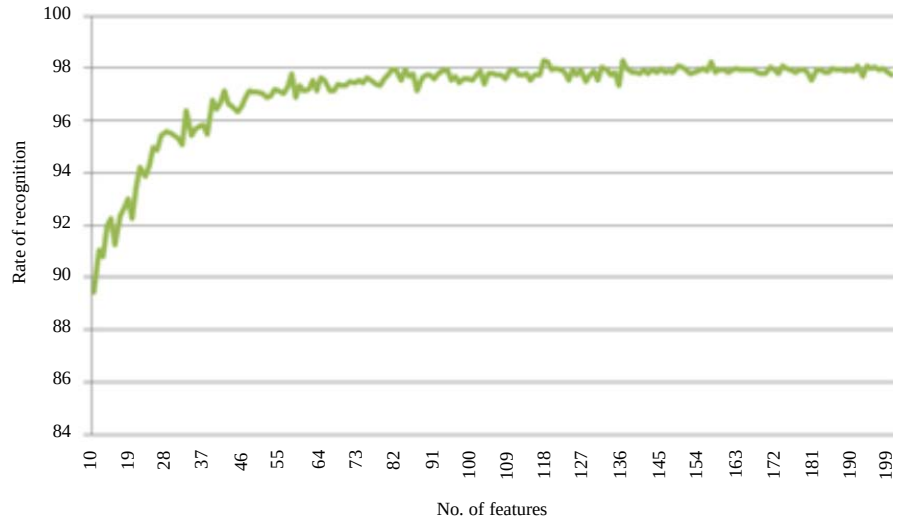

Fig. 6: Recognition rate changes with number of feature change

Table 5: The result of recognition with different distance criteria

Types of data/distances	2500 initial data (%)	2500 test data (%)	5000 data (%)
Euclidean	97.68	97.28	97.48
City block	98.24	97.80	98.02
Hamming	98.24	98.12	98.18

search space about 90.8%. Table 5 shows the results of search spacereduction for each one of limiters and the effect of limiters base on order of usage in the study for first 2500 data and all 5000 test data.

As provided in Table 5 by using all the limiters at the same time, the search space for first 2500 data is limited from 12700 to about 81.2 sub-words with 99.4% reduction and for 5000 test data with 99.36% reduction of search space is limited to 83.7 sub-words.

Using classifier of closest neighbor with zoning and loci feature exist in next stage. In this stage by choosing the first two closest options to input sub-word by using of each of both ultra-search space feature for next stage is limited up to four sub-words. The distance criteria used in this stage accompanies common test criteria and according to the Euclidean distance result.

It is noteworthy that 133 sub-words out of 2500 initial sub-word and 282 out of 5000 sub-word of test data after using search space limiters, the remained search space includes <4 sub-word to which recognition transfer to final recognition stage, directly.

This stage error with only 2 errors in 2500 initial data and 3 errors in a set of 5000 tests equals 0.06%. 2 repeated errors in 2500 initial data which are repeated in 5000 test data resultedthe selected algorithmas if add the number of choice to 4 choices both errors are removed but this causes increasing of next stage choices and consequently increasing final error so it is ignored.

The 3rd error which can be seen in 5000 test data, achieved from search space limitation stage and is due to lack of setting a sub-word in a defined restriction space (in ratio of the number of image top half black dots to image black dots of sub-word thespace limiter due to sweep error).

The last stage of recognition in this study is recognizing sub-word among up to 4 introduced sub-word in previous stage. According to high similarity of limited sub-words, to recognize final sub-word from these 4 sub-words, the features of the number of intersection text field in vertical direction and classifier of the closest neighborhood are used which have well partial view. With respect to sub-words changeable length we have some different features for each sub-word in first step even which with changing the font size, the number of a sub-word changes, too.Therefore according to explained method about how to extract features of the number of vertical intersection text field, features length identification should be done. Regarding the range of sub-words length change, the usage of different features length to reach the best result of recognition was checked. Shows how the rate of recognition changes with feature length change for 2500 initial data; according to result of this diagram, feature length 118 is selected as useful length to continue. Is the result of using this classifier and mentioned feature with three different distance criteria. According to results of this table, Hamming distance criteria are used as the best criteria for this.

The set of sub-words which are applied by using algorithm in the study in 5000 sub-word of test data which are not recognized truly. The used code for size of font-resolution are code 1 for 10 font size scanned with 300 resolution, code 2 for 12 font size scanned with

200 resolution, code 3 for 12 font size scannedwith 300 resolution, code 4 for 14 font size and sweep with 200 resolution and code 5 for 14 font size scanned with 300 resolution.

The sub-word "ﺗﺘﺨﺎ" and "ﺑﻨﻔﺸﻪ" are recognized just once in five statuses. The other 51 sub-words each one are not recognized just in one of the statuses. shows the sub-word which are not recognized correctly and shows the sub-words with result of recognition of these sub-words.

CONCLUSION

Due to the continues structure of Persian writings and expressed faults for methods base onbreaking word to letter, performance of methods base on overall shape is considerable. The large number of classes equals the number of sub-words counts as faults and problems of this way. So, the search space reduction is the first stage to solve the problem.

In this study, according to the goal which is design a simple and useful way for recognition of Farsi typed sub-words, the reduction of search space by simple and useful features is used in the first stage. The reduction of search space is done in several steps. In the first step, search space is just limited to sub-words which the ratio of width to height of them are in a specified restriction, in next levels the same process with the ratio of the number of image black dots to sub-word image surface, the ratio of the number of image top half black dots to image black dots of sub-word and the ratio of the number of image right half black dots to image black dots of sub-words repeat in previous restricted space and finally signs position is used to reduce the search space. In recognition stage, initially by using loci feature and the criteria of the smallest distance, two closer choices to input sub-word are selected. This stage is repeated with zoning features, too. In final stage, recognition of selected sub-word happens from four selected sub-word of previous stage by using of the number of intersection text field in vertical direction and classifier of the closest neighborhood. In designed stages to reduce search space, this space is reduced from 12700-83.7 sub-words with 99.36% reduction (for set of test data) and finally result in the recognition of >98% is achieved for test. The major features of the used method is simplification and at the same time effectiveness of this way.

REFERENCES

Adamek, T., N.E. O'Connor and A.F. Smeaton, 2007. Word matching using single closed contours for indexing handwritten historical documents. Intl. J. Doc. Anal. Recognit., 9: 153-165.

Alibeigi, M., 2017. Recognition of typed Persian sub-words. Masters Thesis, University of Birjand, Birjand, Iran.

Amin, A., 2000. Recognition of printed Arabic text based on global features and decision tree learning techniques. Pattern Recog., 33: 1309-1323.

Bai, S., L. Li and C.L. Tan, 2009. Keyword spotting in document images through word shape coding. Proceedings of the 10th International Conference on Document Analysis and Recognition ICDAR'09, July 26-29, 2009, IEEE, Barcelona, Spain, ISBN:978-1-4244-4500-4, pp: 331-335.

Ebrahimi, A. and E. Kabir, 2008. A pictorial dictionary for printed Farsi subwords. Patt. Recognit. Lett., 29: 656-663.

Ebrahimi, A., 2007. The overall shape of printed sub-words in document image retrieval and recognition Persian texts. Ph.D Thesis, Tarbiat Modarres University, Tehran, Iran.

Fathi, F., 2011. Extraction of index letters from printed Persian sub-words. Masters Thesis, Sahand University of Technology, Tabriz, Iran.

Khosravi, H. and K. Ehsanollah, 2011. Assessment of Farsi litreatures recognition methods based on overall shape of sub-words. Iran. J. Electr. Comput. Eng., 7: 280-267.

Li, L., S.J. Lu and C.L. Tan, 2007. A fast keyword-spotting technique. Proceedings of the Ninth International Conference on Document Analysis and Recognition ICDAR Vol. 1, September 23-27, 2007, IEEE, Parana, Brazil, ISBN:978-0-7695-2822-9, pp: 68-72.

Lorigo, L.M. and V. Govindaraju, 2006. Offline Arabic handwriting recognition: A survey. IEEE Trans. Pattern Anal. Mach. Intell., 28: 712-724.

Lu, Y. and C.L. Tan, 2004. Information retrieval in document image databases. IEEE. Trans. Knowl. Data Eng., 16: 1398-1410.

Madhvanath, S. and V. Govindaraju, 2001. The role of holistic paradigms in handwritten word recognition. IEEE Trans. Pattern, Anal. Machine Intell., 23: 149-164.

Madhvanath, S., G. Kim and V. Govindaraju, 1999. Chaincode contour processing for handwritten word recognition. IEEE. Trans. Patt. Anal. Mach. Intell., 21: 928-932.

Rath, T.M. and R. Manmatha, 2007. Word spotting for historical documents. Intl. J. Doc. Anal. Recognit., 9: 139-152.

Rehman, A. and T. Saba, 2012. Off-line cursive script recognition: Current advances, comparisons and remaining problems. Artif. Intell. Rev., 37: 261-288.

Rodriguez-Serrano, J.A. and F. Perronnin, 2009. Handwritten word-spotting using hidden Markov models and universal vocabularies. Pattern Recognit., 42: 2106-2116.