Elviz A.Ismayilov[1,2] DOI: 10.25045/jpit.v09.i2.11
[1]Azerbaijan State Oil and Industry University
[2]Azerbaijan Republic Customs Committee's Academy, Baku, Azerbaijan
elviz.ismayilov@gmail.com

STUDY OF AZERBAIJANI HAND-PRINTED CHARACTERS RECOGNITION SYSTEM BY NEW FEATURE CLASS AND SVM METHOD`

Although there are widely spread Latin scripts in Azerbaijani alphabet, intending special symbols and morphological content of the language requires individual approach for character recognition. In this paper, "soft" (close to human mind, constructed on base of characteristics used in alphabet learning) features and SVM for recognition of Azerbaijani hand-printed characters are used. For character classification, bootstrap resampling procedure of support vector machines is used. Results are compared with results of other feature classes and methods.

Keywords: *"soft" features, SVM, hand-printed symbols, recognition system.*

Introduction

Artificial neural networks, one of the main trends in artificial intelligence, are widely used in the construction of recognition systems as a classification method. However, it is required to change the structure of the neural network and computational algorithms to solve different recognition problems and increase the accuracy of the systems operation, which complicates the construction of the system. Moreover, the main problem of the use of neural networks is the lack of the same results during the system operation, whereas the desired result is obtained during the learning. In order to solve this problem, Vapnik V.N. proposed a Support Vector Machine (SVM), based on the theory of risk minimization, which incorporates mathematical solutions for learning and recognition compromise [l].

Several methods have been used by different Azerbaijani scientists in different ways (artificial neural networks, theory of fuzzy sets, simple, SVM etc.) for the recognition of print, handwritten and published manuscripts in the Azerbaijani language [2-4]. Whilst using all these methods, different classes of features have been used, and the advantage of each newly created system over another was substantiated by experiments. Nevertheless, the issue of printed handwritten text recognition in the Azerbaijani language has not been fully resolved. Attempts to increase the accuracy of the recognition systems through different methods or new classes of features are of great importance. In this regard, this paper describes a system based on the structural features and bootstrap core SVM for the recognition of print handwritten characters in the Azerbaijani language, and performs experiments to substantiate the results.

A large number of features have been offered by researchers to recognize the characters of different alphabets (especially the alphabets used by ethnic minorities) [5-8]. Nonetheless, most of these features are not understood by human beings, and intended for convenient computer processing and calculations. Therefore, these features are both numerous, and their computing requires considerable time and high performance processing tools. However, despite all this, the quality of the recognition is not at the preferred level, and the system errors become difficult to understand. In this article, the features proposed for the recognition of handwritten texts in Azerbaijani language are easily understood by the human. These features are based on the characteristics a human being uses to remember the letters. Consequently, these features are conditionally called "soft" features. However, it should be admitted that these features are relatively difficult to identify, i.e., they require different approaches and algorithms. In spite of this, the quality of the recognition increases, and the time required for the features' calculation considerably decreases.

Thus, the main advantage of the proposed approach is increased recognition quality by applying more informative class of features, and saving time and resources by reducing the amount of computations. The trial of the proposed structural features in different systems and their comparison with other classes of features is widely explained in the end.

The dictionary approach is used to form a recognition system according to the morphological composition of the Azerbaijani language. A vocabulary containing the Azerbaijani words is drawn up, and the words in this dictionary are accessed to increase the reliability of decision-making during recognition. If there are features that cannot be explicitly identified during the recognition of a word, the most similar word is found in the dictionary, and non-classifiable features are defined based on the dictionary. Thus, the classification of characters also uses a dictionary system containing the Azerbaijani words with certain weights.

Problem statement

SVM defines the optimal separator hyperspace as a space that separates various classes with high difference. Optimal hyperspaces can be determined from the following conditional minimization solution:

$$\begin{aligned} &Min: \frac{1}{2} w^T w, \\ &y_i (w^T x_i + b) \geq 1, i = 1,...,l \end{aligned} \tag{1}$$

where (x_i, y_i) denotes the coordinates of the points, $\frac{b}{\|w\|}$ - equals to the distance module from the coordinate beginning to the hyperspace, w - the normal vector of the hyperspace. For the cases where the lines are inseparable, a modified minimization problem is used by including a new data set $\xi_i, i = 1,...,l$:

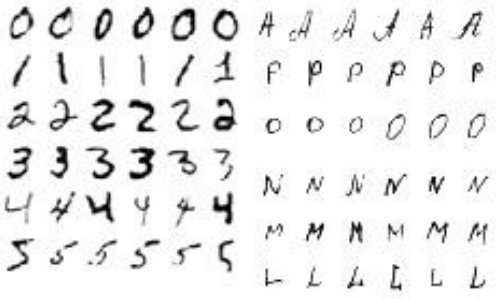
$$\begin{aligned} &Min: \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i, \\ &y_i \left(w^T x_i + b \right) \geq 1 - \xi_i, i = 1,...,l \end{aligned} \tag{2}$$

where C is regulation parameter and defined as $\xi_i = \max\left(0; 1 - y_i f\left(x_i\right)\right), i = 1,...,l$.

The main advantage of the SVM classifier is being non-parametric. As the purpose of the process is the optimization of the separator hyperspace determination, it does not express sensitivity to the statistical distribution of input data. In this regard, it is more convenient to use Bootstrap Resampling (BR) for the selection of the objects' features and for the joint use of SVM classifier [9-10]. This paper uses the BISSP (Backward Input Space Selection Procedure) procedure of BR-SVM method to recognize the printed handwritten Azerbaijani text.

Processing of characters

During the experiments, the samples of printed handwritten characters in the Azerbaijani language written by different persons (100 examples for each of 32 letters and 10 figures, total 4200 characters) are used as a training data (figure 1).



(a) *(b)*

Figure 1. Examples of training data characters: *a)* figures; *b)* letters

Prior to calculating the features, one of the important points is the processing of original document. In other words, selection of the object to be recognized from the general document, clearing the object from the noise, extracting its skeleton and reducing it to the same size (Figure 2). Zhang-Suen algorithm is used for extracting the skeleton of the selected characters after being reduced to the same size [11].

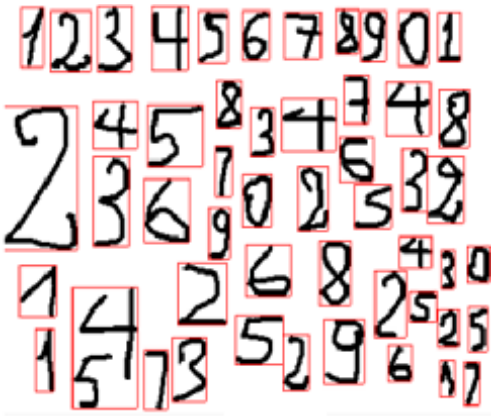


Figure 2. Selection of objects

Extraction of the features

One of the key steps in establishing recognition systems is to generate the features, or more precisely, to evaluate their quality. The features should be chosen with the aim of providing a complete description of the object and not be close to each other. The AdDel Algorithm is used to identify the vectors containing more informative features for the given object [12]. As a result, the following features are informatively selected for the recognition of the print handwritings in Azerbaijani language:

A. The number of straight lines drawn to the rectangle where the character is located

After being processed, inclined lines are drawn horizontally, vertically and horizontally at different angles from the border of the rectangle, and the number of points, where the characters are crossing with the lines, are defined (Figure 3). These lines are defined experimentally, i.e., multiple lines are drawn from the border of the rectangle where the characters are located and the coordinates of the starting and end points are randomly selected, and the experiments are carried out. As a result, six straight lines that classify symbol classes with higher precision are maintained, and the number of their intersection points with the characters is included in the feature vector.

Figure 3. Determination of intersection points

As a result of the optimization of the feature vectors, the number of intersection points of the straight lines crossing the following points out of the straight lines passing through different points was found to be more favorable for the classification of print handwritings in Azerbaijani (Table 1, where $(x_0; y_0)$ are the coordinates of starting point, $(x_1; y_1)$ - of end point).

Table 1.

Coordinates of the points through which the straight lines pass

N	**1**	**2**	**3**	**4**	**5**	**6**
$(x_0; y_0)$	(0; 20)	(0; 40)	(21; 0)	(0; 50)	(0; 50)	(0; 50)
$(x_1; y_1)$	(42; 20)	(42; 40)	(21; 63)	(42; 40)	(42; 25)	(42; 10)

B. Closed area of characters

Closed areas must be identified when classifying the characters. The following algorithm is used to determine the closed area. The first colorless pixel of the character is identified and colored with the neighboring colorless character, and then the other colorless pixels are identified and colored together with neighboring pixels to other color. As a result, if two colors are obtained, the character has a closed area (another color determines the background of the character), whereas three colors indicates the presence of two closed areas, and so forth (Figure 4).

Figure 4. Identification of closed area

It should be noted that wrong information may also be obtained after the skeleton extraction, for example, in the case of letter "o" or figure "0", sometimes two closed areas are obtained, however, in fact, there should be one closed area (figure 5). In order to avoid such problems, the diagonal of the areas is determined., The diagonal of the largest area is taken as a feature.

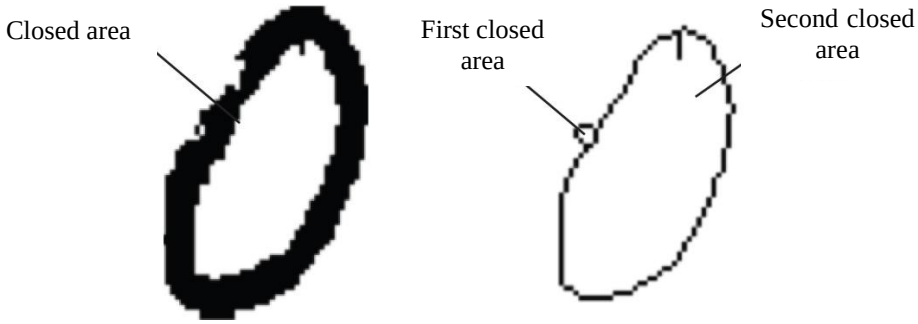


Figure 5. Closed areas before and after skeleton

C. Location of the closed area

Location of the given character is located on the center of gravity. Then the distance from the center of gravity of the area to the lower border of the rectangle is calculated, which is very useful for distinguishing certain class of characters. For example, the largest closed area of the letter "p" is the upper part, while this area of the closed letter "b" is in the lower part.

Thus, the proposed "soft" features vector contains 8 elements:

$F = \{f_1, \ldots, f_6$ – the number of crossing points, f_7 - diagonal of the largest closed area, f_8 - distance from the center of gravity of the closed area$\}$.

Experimental results

In order to substantiate the superiority of the class of features proposed in the article, the following features applied by various researchers for the recognition of the print handwritings in

Azerbaijani: the results are comparatively analyzed.

Features of class I. To recognize the print handwritings in the Azerbaijani language, the characters are normalized to the size of 28x28 through Ayda-zadeh and Mustafayev and stored in gray shades in the database. The color value (RGB) of all the pixels in the rectangle, where the character is located, is taken as a feature. Hence, features vector 784 (28x28 = 784) ranged from 0 to 255 is calculated for each sample [2].

When calculating the **features of class II**, the character is normalized to 16x24 pixels. The normalized image is divided into 22 sections in 4 directions - 1 vertical, 1 horizontal and 2 diagonals (4 vertical, 6 horizontal and for each diagonal). 4x4 pixel square particles with at least one black and one white pixel (not quite black and white) are taken from each section. All possible variants of these square particles by pixels are 14, and the direction is determined depending on the location of white and black pixels in these variants. Thus, 4 directions are available in these 14 variants: vertical, horizontal, left-to-right diagonal and right-to-left diagonal. At the end of the algorithm, the number of 4 directions mentioned in the latter 22 particles is calculated, and consequently, the contour direction vector contained 22x4 = 88 element, which defines the direction of the image by its contour, is obtained. 88 features are obtained through this method [3].

Features of class III. These features include a set of features identified by the Peripheral Directional Contributivty (PDC) algorithm [4]. Each point of the character is a 8- or 4-dimensional vector. Each component of the vector holds the distance to the border of the character and square. When determining the features of the characters, 8 directions are taken, and the crossing point of straight lines connecting them is defined. This point is called the first level linear point (LP). The second LP, including higher-level LPs are found similarly. The size of the PDC algorithm depends on the condition of the task. The parameters of the PDC method for features of print handwritten characters in Azerbaijani are constructed as follows: DC size - 8, number of revisions - 4, depth – 1, and number of segments - 8. The total number of PDC features within these parameters is 8x4x1x8 = 256.

The provided instructional samples are identified using "soft" features and each of the features from all 3 classes described above through SVM-s bootstrap core and artificial neural networks. The results of the recognition are provided in Table 2:

Table 2.

The results of recognition through SVM and artificial neural networks

	"Soft" features	Class I	Class II	Class III
Number of features	8	784	88	256
SVM Bootstrap	93,80 %	90,05 %	88,06 %	79,03 %
Artificial neural networks	86,22 %	80,31 %	86,10 %	82.47%

Table 2 shows that as the character of the features changes, the result of recognition also varies; the best result is performed through the application of the "soft" features and SVM method proposed by the author.

During the experiment, 60% of the training base is used for teaching, 20% - for assessment of learning and 20% - for the system settings. The best performance was obtained during recognition with the SVM Bootstrap, however; in this case, the problems with the recognition of certain letters occur. Table 3 depicts incorrectly recognized characters and their alternate characters.

Table 3.

Incorrectly recognized characters

Experimented characters	d	ə	k	z	q
Incorrectly suggested characters	o, a	z, b	r,y	x, f	e, o

In order to research the inaccuracy, incorrectly recognized characters are re-trained separately and the verification is repeated. As a result, although, only distinguishing the letters z and x creates some problems in the system, other characters are recognized by the system at a high level.

Conclusion

The results of numerous experiments performed on various features showed that SVM could be used successfully not only for the recognition of print handwritings, but also for building other recognition systems. The results of the research in the article were summarized as follows:

1) Whilst defeatureing a recognition system, it was possible to define which class of features was more useful for a given issue through "mutual recognition" method with the use of the group of features;
2) Moreover, different results could also be achieved by changing the new features cores and allowing variations in the class of features. Although the number of features representing the classes affected the time spent on learning, it did not affect the accuracy of the recognition, thus, the basic requirement for the features was that they were informative and most unique for each class.

References

1. Cristianini N., Shawe-Taylor J. An Introduction to support vector machines and other kernelbased learning methods // Cambridge University Press, 2000.
2. Aida-zade K.R., Hasanov J.Z. Cursive Handwritten Azerbaijani Latin Text Segmentation Based on Word baseline / INISTA, Trabzon, Turkey, 2009, pp.63–66.
3. Aida-zade K.R., Hasanov J.Z. Word base line detection in handwritten text recognition systems // International journal of computer systems science and engineering, 2009, no.4, pp.49–53.
4. Aida-zade K.R., Mustafayev E.E. On a hierarchical handwritten forms recognition system on the basis of the neural network / Proceed. Inter. Conf. TAINN, Canakkale, 2003.
5. Moubtahij H.E., Halli A., Satori K. Review of feature extraction techniques for offline handwriting arabic text recognition // International journal of advances in engineering & technology, 2014, pp.50–58.
6. Hadidi G., Delavari H. Persian handwritten words detection based on features, extraction and fuzzy algorithm // Electrical and Electronics Engineering: An International Journal (ELELIJ), 2015, vol.4, no.2, pp.93–104.
7. Hussain E., Hannan A., Kashyap K. A zoning based feature extraction method for recognition of handwritten assamese characters // International journal of computer science and technology, 2015, vol.6, no.2, pp.226–228.
8. Lawgali A., Bouridane A., Angelova M., Ghassemlooy Z. Handwritten arabic character recognition: which feature extraction method? // International journal of advanced science and technology, 2011, vol.34, pp.1–8.
9. Platt J. Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods — Support Vector Learning // MIT Press, 1999, pp.185–208.
10. Atienza F.A. Bootstrap feature selection in Support Vector Machines for ventricular fibrillation detection / ESANN'2006, Belgium, 2006, pp.233–238.
11. Chen W., Sui L., Xu Z., Lang Y. Improved Zhang-Suen thinning algorithm in binary line drawing applications/ICSAI, 2012, China, pp.1947–1950, doi: 10.1109IICSAI.2012.6223430
12. Ismayilov E., Ismayilova N. Fuzzy Features Extraction for Hand-printed character/digit recognition system / INISTA, Italy, 2014, pp.249–253.