DOI: 10.21917/ijivp.2018.0265

APPLICATION OF SVM AND SOFT FEATURES TO AZERBAIJANI TEXT RECOGNITION

Elviz A. Ismayilov

Department of General and Applied Mathematics, Azerbaijan State Oil and Industry University, Azerbaijan

Abstract

The purpose of this study is to establish more accurate and less time-consuming recognition system for Azerbaijani text recognition. The main problem of investigating and developing recognition systems is the extraction of features, in view of the fact that, most of current recognition systems use features, which are unintelligible for human mind and proposed for operating by computers. For eliminating above-mentioned problem, in this paper was offered "soft" features, extracted on base of human-mind techniques. On the side of validating SVM approach and "soft" features provided in this paper, experiments were executed using various feature classes offered for Azerbaijani hand-printed characters and different methods.

Keywords:

SVM, Soft Features, Hand-Printed Characters, Characters Recognition, Features Extraction

1. INTRODUCTION

Many problems need to be solved in order to read text from different sources [1]-[3]. Text recognition became an important task in computer intelligence and image processing for the reason of its applications in image understanding, visual facilitation and etc. [4].

Today many researchers offer different techniques and methods for text reading. Campos's and others application of nearest neighbor and SVM classification method is extremely useful, because it sheds light on the difficult problem of character recognition from texts [5].

Many approaches based on artificial neural networks were created for text reading and character recognition [6] [7]. The main disadvantage of created systems is their usability only for a few languages, therefore the main part of world languages is in excess of present text recognition techniques.

A lot of issues require solution in the interest of character recognition covering image processing, text segmentation, character thinning, feature extraction and etc. Our goal in this paper is creating more reliable recognition system for Azerbaijani hand-printed character and handwritten text recognition.

In Azerbaijan, researchers applied different methods (artificial neural networks, fuzzy sets theory, SVM, etc.) to recognition of Azerbaijani printed, hand-printed and handwritten texts [8]-[10]. They used different feature classes during application of designated methods; superiority of each new system was substantiated by experiments.

However, we cannot consider Azerbaijani hand-printed recognition task as 'ended' problem. Each effort for increasing of recognition using different methods and new feature classes are significant. For this purpose, in the paper was described hand-printed character recognition system created on base of pattern features and bootstrap resampling SVM method. Usefulness of offered system is approved by experiments.

As mentioned by numerous scholars feature extraction is one of the three main stages in developing recognition systems. Researchers offered various feature classes for hand-printed character recognition as well as handwritten text recognition of different alphabets (particularly alphabets used by minor nations) [11]-[15]. Nevertheless, main part of these features is incomprehensible for human mind and intended for processing and calculation by computer. Therefore, quantity of these features is large and their calculation requires more time and high performance.

In spite of high number of features, quality of recognition is not on desirable level and it is difficult to define errors source of the system. Features offered in this paper for Azerbaijani hand-printed characters are simple and close to human mind. Thus, these features have been extracted by the characteristics, which we use on the alphabet learning process. Literally we'll call them "soft" features. Let's admit that, extraction of such features is difficult, as requires different approaches and algorithms. Though these disadvantages, application of "soft" features increase accuracy of recognition and decrease the time required for features calculation.

Main superiority of offered approach is application of the most informative feature class for classification. As a result, the quality of recognition increases and provides economy of time and resources by decreasing volume of calculations. At the end of the paper were interpreted testing results of offered features in different systems and their comparison with results of other feature classes.

As has been considered the main goal of text recognition systems is not only classification of characters, but finding of the whole words [16]. Created recognition system is substantiated by dictionary consisting of words in Azerbaijani language, recognition system addresses to this dictionary for increasing reliability of decision-making. Thus, if an undefined symbol exists in recognizing word, then system finds the word with the highest matching from the dictionary. Hence, in classification process dictionary composed by Azerbaijani words is used with particular weight.

As cited above most of approaches offered for Azerbaijani character recognition based on neural networks. In this study we use Support Vector Machine (SVM) for recognition, since SVM method has significant advantages as reaching global optimum by quadratic programming, requiring less memory to catch the prognostic model and etc.

2. MAIN APPROACH

SVM method is one of the best tools applied for learning models with data analyses and classification problems. The priority of SVM classifier has been confirmed in different recognition systems as well as on recognition of Gurmukhi characters and handwritten Kannada characters [17] [18]. In Support Vectors Machine (SVM) method optimum separating hyper plane is plane which separates different classes with the greatest difference [19]. We can define optimum hyper plane as solution of following conditional minimization task:

$$Min:\frac{1}{2}w^T w, \tag{1}$$

$$y_i\left(w^T x_i + b\right) \geq 1, i = 1,...,l \tag{2}$$

In cases where data set is not linear separable, we can use modified minimization task:

$$Min:\frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i, \tag{3}$$

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, i = 1,...,l \tag{4}$$

Here C is the regulation parameter, and $\xi = \max\left(0;1 - y_i f\left(x_i\right)\right)_i, i = 1,...,l$.

The main goal of SVM classifier is being parameter less method. Aim of the method is optimization of hyper plane definition task and therefore method does not demonstrate sensitivity to statistical distribution of input variables. In accordance with this position, application of Bootstrap Resampling (BR) method is convenient for using of SVM classifier with features selection of recognizing objects [20] [21]. In this paper BISSP (Backward Input Space Selection Procedure) procedure of BR-SVM was used for recognition of Azerbaijani hand-printed character.

3. IMAGE PROCESSING AND FEATURE EXTRACTION

Recognition database expressed by Azerbaijani hand-printed character examples, digits and special symbols (1000 examples for each class of 32 letters, 10 digits and special symbols – totally 52000 characters) (Fig.1).

The first step in character recognition process is recognizing object selection from whole image, noise cleaning, thinning and fitting all symbols to the same size. In this paper, Zhan-Zuen method for thinning was used [22].

Fig.1. Examples of characters from training database

In creating recognition systems, the main procedure is feature extraction procedure, exactly quality level evaluation of features. During forming, the system must have selected those features, which more precisely characterizes symbol classes. For determining the most informative feature class for recognizing class was used AdDel algorithm.

Consequently, for recognition of Azerbaijani hand-printed characters were selected following informative features.

4. "SOFT" FEATURES

4.1 NUMBER OF INTERSECTION POINTS

With lines drew into rectangle, in which recognizing symbol has been fitted. We drew horizontal, vertical and slope lines with different angles from bound of the rectangle after processing and determine number of intersection points of these lines with symbol (Fig.2). More useful 6 lines have been chosen experimentally, among lines with randomly selected start and end points (Table.1);

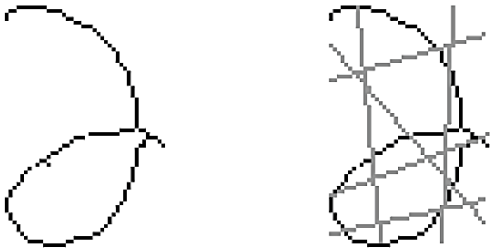

Fig.2. Determination of number of intersection points

Table.1. Start and end points of lines added to features vector

N	1	2	3	4	5	6
Start points of the drawn lines	(0;20)	(0;40)	(21;0)	(0;50)	(0;50)	(0;50)
End points of the drawn lines	(42;20)	(42;40)	(21;63)	(42;40)	(42;25)	(42;10)

4.2 CLOSED AREA

The character is one of the most important parameters of symbols as "O", "P", "B", etc. Such properties of closed area in symbol as quantity, size and location are useful in recognition of these characters.

For determining the quantity of closed areas in the symbol following algorithm was used: define the first colorless pixel, paint it with the neighbor colorless pixel until there is no colorless neighbor pixels. Then define another colorless pixel and paint it with neighbor pixels to different color. Getting two colors at the end corresponds to one closed area, three colors to two closed areas, and etc. (Fig.3).

Fig.3. Fixing of number of closed areas

Let's note that, it is possible to get falsehood closed areas after thinning, in Fig.4 after thinning of the symbol "O" we got two closed areas. To implement this ambiguity, we determine diameters of closed areas, and find ratio between them, closed areas with small diameters counts as non-significant closed areas;

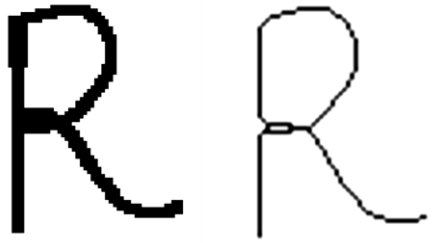

Fig.4. Discovery of falsehood closed areas

Location of the largest closed area is defined by gravity center. We calculated distance among gravity center of the symbol and lower bound of rectangle. This feature is very useful for comparing symbols as "P" (where the largest closed area located above) and "Ɵ" (where the largest closed area located below).

Thus, offered "soft" features class consists of 8 elements: {F_1, F_2, …, F_8}, where F_1, …, F_6 number of intersection points, F_7 – number of significant closed areas, F8 – location of the largest closed area.

5. EXPERIMENTAL RESULTS

In this section we will describe results of the experiments carried out for Azerbaijani hand-printed character recognition system with "soft" features and SVM method, its comparative study with other systems based on different features and neural networks. We have selected 3 different feature classes, all of which were offered for Azerbaijani hand-printed character recognition:

- *I class*: In the first class of features offered by Ayda-zade and Mustafayev, symbols normalized to 28 x 28 and stored in database as shades of grey. In this method features defined as color value of all pixels in rectangle. Thus, for each example features vector consists of 784 elements, which get values from 0 to 255;
- *II class*: For determination of the second class of features researchers used following algorithm. Character normalized into rectangle 16×24, then normalized image was separated into 22 parts by 1 vertical, 1 horizontal and 2 diagonal lines. In each part were selected squares 4×4, which has at least one white or one black pixel. Total number of these squares does not exceed 14. Then was defined direction on base of position black and white pixels. At the end of this algorithm was defined contour direction vector with 22×4 = 88 elements;
- *III class*: Features calculated by Peripheral Directional Contributivity (PDC) were analyzed as the third class of features. Each point of the symbol can be characterized as a vector with 8 or 4 elements. Components of vector determine distance among symbol and bound of the square. Intersection point of lines connected 8 directions was defined and denoted as the first grade linear point, then was defined the second grade linear point, and etc. The size of PDC algorithm depends on task condition.

For Azerbaijani hand-printed characters, parameters for PDC are: size of DC - 8, number of directions - 1, depth - 1, number of segments - 8. So, the total number of PDC features is 8×4×1×8=256;

Recognition results were compared for SVM and neural networks method on base of listed above feature classes (Table.2). According to results we can see that, quality of recognition changes for different features. The best results were obtained in case of "soft" features and SVM method combination. Let's note that, by this approach was obtained 98.6% recognition rate for hand-printed digits.

Table.2. Recognition results of experiments

Features	Recognition Rate					
	Letters		Digits		Special Characters	
	by ANN	by SVM	by ANN	by SVM	by ANN	by SVM
"Soft" features	86.22%	95.40%	93.23%	98.60%	91.52%	96.20%
I class of features	80.31%	90.05%	87.88%	89.93%	83.22%	87.26%
II class of features	88.06%	80.10%	93.27%	85.54%	89.03%	80.37%
III class of features	79.03%	82.47%	87.44%	89.07%	81.64%	89.06%

6. CONCLUSIONS

Results of numerous experiments on base of different feature classes show that, SVM method can successively applied in developing of different recognition systems. Results of current research can be evaluated by following propositions:

On developing of recognition system it is possible to define more informative feature class by using different classes and "mutual comparing" method;

Also it is likely to obtain different results by changing core of feature classes. Although number of features affects training time, it does not influence the recognition accuracy. As a consequence,

main necessity for feature is being informative and maximally unique for each class.

7. DISCUSSION

The main goal of this study was demonstration of priority using "soft" features in recognition systems. Most of research works in this area are devoted to application of features constructed for computer processing and calculating. This approach does not answer to the following question: why and for which objects we use current feature? In contrast to this proposal, "soft" features offered in this paper make possible solution of the problem noted above.

Results obtained from numerous experiments proved usability and importance of features created on base of human brain skills. The highest recognition rate of Azerbaijani hand-printed characters, digits and special symbols has guaranteed profitably application of similar features in different recognition tasks. Although determination of "soft" features requires individual approach and algorithm development skills, their implementation leads to time and resource savings.

REFERENCES

[1] Q. Ye and D. Doermann, "Text Detection and Recognition in Imagery: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 7, pp. 1480-1500, 2015.

[2] Y. Cong and L. Wenyu, "A Unified Framework for Multi oriented Text Detection and Recognition", *IEEE Transactions on Image Processing*, Vol. 23, No. 11, pp. 4737-4749, 2014.

[3] Z. Yingying, Y. Cong and B. Xiang, "Scene Text Detection and Recognition: Recent Advances and Future Trends", *Frontiers on Computer Science*, Vol. 10, No. 1, pp. 19-36, 2016.

[4] Shivani and Dipti Bansal, "Techniques of Text Detection and Recognition: A Survey", *International Journal of Emerging Research in Management and Technology*, Vol. 6, No. 6, pp. 83-87, 2017.

[5] Bodla Rakesh Babu and ManikVarma, "Character Recognition in Natural Images", *Proceedings of 4th International Conference on Computer Vision Theory and Applications*, pp. 5-8, 2009.

[6] Sonia Yousfi et al., "Arabic Text Detection in Videos using Neural and Boosting-based Approaches: Application to Video Indexing", *Proceedings of 4th International Conference on Image Processing*, pp. 1131-1137, 2014.

[7] Weilin Huang, Tong He, Yu Qiao and Jian Yao, "Text-Attentional Convolutional Neural Network for Scene Text Detection", *IEEE Transactions on Image Processing*, Vol. 25, No. 6, pp. 2529-2541, 2016.

[8] K.R. Aida-Zade and J.Z. Hasanov, "Cursive Handwritten Azerbaijani Latin Text Segmentation Based on Word Baseline", *Proceedings of International Symposium on Innovations in Intelligent SysTems and Applications*, Trabzon, Turkey, pp. 63-66, 2009.

[9] K.R. Aida-Zade and J.Z. Hasanov, "Word Base Line Detection in Handwritten Text Recognition Systems", *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, Vol. 3, No. 4, pp. 1133-1137, 2009.

[10] K.R. Aida-Zade, E. Mustafaev and J.Z. Hasanov, "Intelligent Reading System based on Mobile Platform", *Proceedings of International Conference on Problems of Cybernetics and Informatics*, pp. 12-14, 2012.

[11] I. Ahmad et al., "Improvements in Sub-Character HMM Model Based Arabic Text Recognition", *Proceedings of International Conference on Frontiers in Handwriting Recognition*, pp. 14-18, 2014.

[12] Neha V Sharma, P. Kavita and Gaurav Aggarwal, "Offline Segmentation and Script Recognition of Hindi using Knowledge based Approach and Multi Layered Perceptron Neural Network, *Journal of Statistics and Management Systems*, Vol. 20, No. 4, pp. 499-506, 2017.

[13] V. Chavan, A. Malage, K. Mehrotra and M.K. Gupta, "Printed Text Recognition using BLSTM and MDLSTM for Indian Languages", *Proceedings of International Conference on Image Information Processing*, pp. 1-6, 2017.

[14] S. Gaur, S. Sonkar and P.P. Roy, "Generation of Synthetic Training data for Handwritten Indic Script Recognition", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 491-495, 2015.

[15] A. Sethy et al., "Off-line Odia Handwritten Character Recognition: A Hybrid Approach", *Proceedings of International Conference on Computational Signal Processing and Analysis*, pp. 1-7, 2018.

[16] A. Lawgali, "A Survey on Arabic Character Recognition, International Journal of Signal Processing", *Image Processing and Pattern Recognition*, Vol. 8, No. 2, pp. 401-426, 2015.

[17] D. Deepika Sood, "Basic Process of Hand written Gurmukhi Character Recognition: Detailed Review", *International Journal of Current Trends in Science and Technology*, Vol. 7, No. 12, pp. 20561-20569, 2017.

[18] S.A. Angadi and S.H. Angadi, "Structural Features for Recognition of Hand Written Kannada Character Based on SVM", *International Journal of Computer Science, Engineering and Information Technology*, Vol. 5, No. 2, pp. 25-32, 2015.

[19] R.G. Negri, L.V. Dutra and S.J.S. Sant Anna, "Comparing Support Vector Machine Contextual Approaches for Urban Area Classification", *Remote Sensing Letters*, Vol. 7, No. 5, pp. 485-494, 2016.

[20] Girish Chandrashekar and Ferat Sahin, "A Survey on Feature Selection Methods", *Computers and Electrical Engineering*, Vol. 40, No. 1, pp. 16-28, 2014.

[21] B. Bischl et al., "MLR: Machine Learning in R", *Journal of Machine Learning Research*, Vol. 17, No. 1, pp. 1-5, 2016.

[22] Made Sudarma, Ni Putu Sutramiani, "The Thinning Zhang-Suen Application Method in the Image of Balinese Scripts on the Papyrus", *International Journal of Computer Applications*, Vol. 91, No. 1, pp. 9-13, 2014.