VERIZONA PUBLISHER
Publishing Your Ideas

www.verizonaonlinepublishing.com
Vol: 1, Issue: 1

Journal of Electronics and Communication Engineering Research

Machine-Written Character Recognition Using A Supervised Machine Learning Approach

Ehsan Shirzadi

IEEE member, KL, Malaysia

***Corresponding author:** EhsanShirzadi, IEEE member, KL, Malaysia; E mail: ehsan.shirzadi.1984@ieee.org

Article Type: Research, **Submission Date**: 07 March 2016, **Accepted Date:** 15 March 2016, **Published Date:** 04 April 2016.

Citation: Ehsan Shirzadi (2016) Machine-Written Character Recognition Using A Supervised Machine Learning Approach. J. Elec. Commu. Eng. Resol 1(1): 6-10.

Abstract

Machine-Written character recognition is one among the abilities of computers to recognize input characters from external sources such as e-forms, e-letters, and e-documents. The field of character recognition can make appropriate ways to regenerate such a text format from official electronic documents, providing textual information for different enterprise class applications such as G2G, B2B, and B2C. Machine learning algorithms expose a great ability to intelligible interpret the input machine-written characters to digitized characters which could be stored in a particular le for further processing purposes. The objective of this research project is to implement an offline machine-written character recognition system for lower case English characters. I show that a machine-written character recognition system could be implemented using LVQ algorithm as a supervised classification strategy. Several experiments have performed on synthetic data to validate the speed and accuracy of the proposed method.

Introduction

Character recognition has been the subject of considerable researches in the field of pattern recognition and image analysis due to its importance for various applications involving scanning and recognition. These applications include a wide range from invoice, check or legal billing documents reading, to real-time translation of foreign-language signs or creating a reading machine for the blind.

Over the past few years, automatic character recognition has found significant interest from the variety of scientific researches. This is obviously evident from Figure 1 that shows the journal and conference papers related to the rubric which have been published in Elsevier from 2009 to 2014. Machine-written or hand-written character recognition can be fulfilled the following objectives:

- To reproduce textual information from electronic or non-electronic documents.
- To provide a capability to separate machine-written and hand-written within o cial documents.

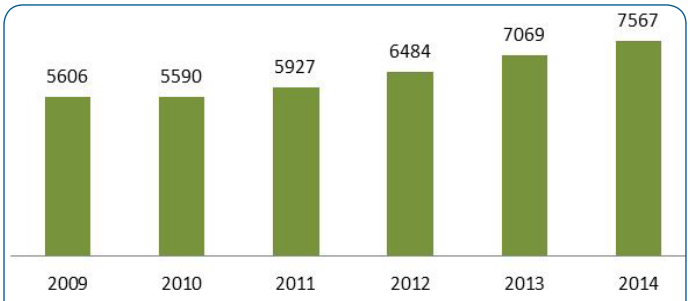

Figure 1: Number of Elsevier publications over last six years. Results obtained by submitting query "character recognition" from Elsevier website at http://www.sciencedirect.com

- To provide a realistic text le from Portable Document Formats (PDF) which stores such a graphical shape model of a lexical character.

Learning Vector Quantization (LVQ) is a well-known supervised classification technique which can be applied to multi-level classification problems in a natural way [1-5]. A key concept in a LVQ algorithm is the selection of an appropriate measure of a similarity for training and classification purpose [6]. The motivation of this work is to use LVQ ability to develop an intelligent platform for machine-written character recognition. In this research project, I implement a supervised classification strategy based on the LVQ algorithm for o ine machine-written system to recognize lower case English characters. The contributions of the work are as following: 1) De ne and describe an encoding system for each lower case English character separately, 2) Design and implement an offline machine-written system based on the LVQ neural networks, and 3) Test and verify the proposed algorithm for each lower case character.

The rest of the paper is arranged as follows. I first give a brief literature review of both character recognition and LVQ classification technique in Section 2. In Section 3, I explain the proposed method including an encoding system and the proposed LVQ algorithm. I present a brief demo of the Matlab implementation in Section 4. The experimental validations performing the proposed model on synthetic data are shown in

Section 5. Conclusion and possible future direction are presented in Section 6.

Related Work

I start by a brief literature review to reflect the current knowledge and recent advances in character recognition techniques and LVQ classification strategy.

Character Recognition

Two main domains of character recognition are: Off-Line and On-Line character recognition. In On-Line character recognition system, document is first generated, then captured optically by a scanner, digitized and stored in computer and finally taken for testing purpose and processing. In contrast to the O -Line class, in the On-Line systems, character is processed while it was under creation. So each point of the pattern is a function of time, pressure, speed, slant, strokes and etc. [7,8].

In 2012, Dash [9] proposed an Associative Memory Net (AMN) based O - Line character recognition system, implemented with C programming language. AMN is a neural network which can store patterns as memories. He also developed a parallel algorithm using an API package OpenMP (Open Multiprocessing), to improve the system performance and make it time efficient. The characters collected from system are used to train the AMN and characters collected from different persons are used for testing the recognition ability of the system. Extraction of pixels from the characters is done followed by implementation of auto AMN for both training and testing. When the network is being tested with a key pattern, it corresponds by producing one of the stored patterns, which closely resembles to key pattern. This method reported an accuracy rate of 72.20% in average and 88.5% in the best case.

In 2014, Mahajan et al. [10] proposed and demonstrated an opticalcorrelator neural network architecture for character recognition. The proposed solution forms the basis of an OCR which is trained using the Back Propagation algorithm where binary numbers are used for representing each typed English character. A feature extraction system takes these binary numbers as input. Then the output of this system and the input are fed to the ANN. After the Feed Forward Algorithm, the Back Propagation Algorithm performs training, calculating error, and modifying weights. Various kinds of changes can be pro-posed in feature extraction to improve this method. One which has been done in this paper is the increasing of the database used for training the ANN, so as to enable it to recognize stylized fonts and hence demonstrated the capabilities of artificial neural network (Back Propagation network) implementation in recognition of characters. In this method, the input to the system can be in the form of hand-written or typed document. In case of typed document, some preprocessing techniques like noise removing and scaling is performed. Next step is segmentation. After segmentation of the given string, the features are extracted. Finally, the extracted feature is given as an input to the Neural Network. The output of the ANN will be the recognized character. the authors also explained how the working of the OCR creates an image le for the characters in these steps: First, they close the image to get rid of minor holes. Then resize the image to 16 x16 matrices and thin the image so only skeleton remains. Finally, they determine all the histograms horizontal, vertical, right diagonal and left diagonal. When the concatenation of the histograms is done the result for an image is obtained.

In 2014, Mori et al. [7] recommended the use of global features for On-Line character recognition. Most On-line character recognition methods have usedonly local features such as the XY-coordinate feature and the local direction feature to represent character strokes. Global features represent the relationship between two temporally distant points in a handwriting pattern. In this case, they defined it as the relative vector between two temporally separated points. This relationship cannot be represented by local features defined at individual points. Feature extraction is one of the most important parts of the character recognition. In this case, global feature of character strokes plays an important role in classification accuracy. They employed AdaBoost for feature selection, which is an iterative learning scheme that iteratively selects a classifier, called a weak learner. They performed two sets of experiments. At their first experiment, they used the AdaBoost-based machine learning framework. Results revealed that global features were more important in terms of discrimination power than local features. Then with the recognition experiment they proved that global features yield better classification accuracy not only for training samples but also for test ones.

Readers interested in character recognition and its algorithms and solutions are refereed to [11] for further information.

LVQ

The Learning Vector Quantization (LVQ) algorithm belongs to the area of Ma-chine Learning and Artificial Neural Networks in which it could be employed for supervised classification problems [1].

In 2011, wiercz [12] deployed a novel method based on the wavelet decomposition and the LVQ algorithm to automatic classification of signals with linear frequency modulation, generated by radar emitters. The LVQ algorithm with a previously defined set of features as an input of the LVQ neural network was proposed as the intelligent classification algorithm.

In 2012, Malakooti et al. [13] presented an efficient strategy, based on the vector quantization techniques for the detection of human cells in electron microscopy images. They performed an edge detection method to specify the desired region of any object in image and then applied vector quantization technique to cluster the property approximation of human cells. Their proposed algorithm does not require any under image segmentation. A kind of similar techniques have been applied in [14] to metacarpal bones localization in X-ray Imagery.

In 2014, Melin et al. [15] developed an LVQ based algorithm for classification of electrocardiogram signals (ECG). For this purpose they employed a particular database with 15 classes. Comparing with other approaches with the same database, their proposed system produces very good results. They have shown that working only with LVQ networks succeeded in obtaining better results for arrhythmia classification. They concluded that the proposed modular LVQ neural network is a good approach in solving complex classification problems.

Further studies on LVQ algorithm and its applications can be found in [6,1].

System Design

In this section I further discuss the proposed system to recognize machine-written characters in English language.

For the purpose of specifying English characters, I first proposed an encoding system. By considering the Times New Roman or Arial font styles, I divide each lower case character into 12 blocks or frames (Figure 2A) in which I only have 10 meaningful frames (Figure 2B).

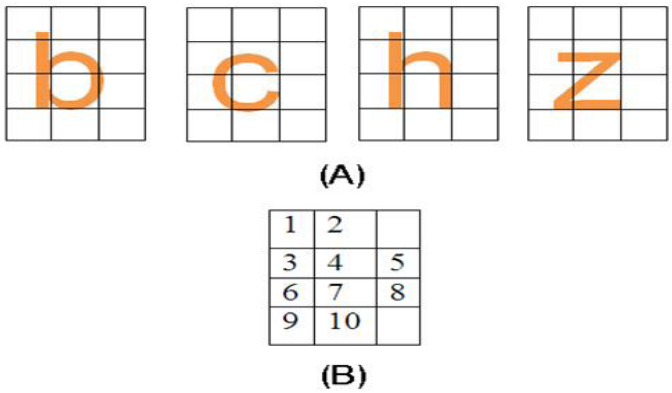

Figure 2: I de ne 12 frames for each lower case character. The frames structure for b, c, h, and z is shown in (A). Frames identification numbers are presented in (B). Lower case English characters never cover the last columns in the firrst and last rows ([1,3], [4,3]) of the proposed frames, so I only have 10 meaningful frames

I then consider a 12 bit sequence to encode each individual frame. After that, I plug the encoding string into LVQ networks. As I said earlier, for each frame in a character, I design a 12 bit sequence. These sequences are defines based on the extensive investigations of each English alphabet using those font styles. I de ne the proposed 12 bit sequence as follows:

Bit 1: If the particular frame has such a vertical line shape, then the bit defines as 1, otherwise 0.

Bit 2: If the particular frame has such a horizontal line shape, then the bit defines as 1, otherwise 0.

Bit 3: If the particular frame has such a skew line shape with about 45 degree with respect to X axis, then the bit defines as 1, otherwise 0.

Bit 4: If the particular frame has such a skew line shape with about 135 degree with respect to X axis, then the bit defines as 1, otherwise 0.

Bit 5: If there is a point in the particular frame (i.e. i or j), then the bit defines as 1, otherwise 0.

Bit 6: If there is a cross section in the particular frame, then the bit defines as 1, otherwise 0.

Bit 7: If a character begins or ends in the particular frame, then the bit defines as 1, otherwise 0.

Bit 8: If there is a similar ∩ shape in the particular frame, then the bit defines 1, otherwise 0.

Bit 9: If there is a similar U shape in the particular frame, then the bit defines as 1, otherwise 0.

Bit 10: If just a portion of the particular frame has a shape like C, then the bit defines as 1, otherwise 0.

Bit 11: If just a portion of the particular frame has a shape like reverse C, then the bit defines as 1, otherwise 0.

Bit 12: If just a portion of the particular frame has a shape like o, then the bit defines as 1, otherwise 0.

For instance, using this 12 bit sequence, Characters "a" and "z" could be presented by 10 frames each includes 12 bit as Figure 3, and Figure 4 respectively.

Frame #1: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #2: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #3: 0 0 1 0 0 0 0 0 0 0 0 0
Frame #4: 1 0 0 1 0 1 1 0 0 0 0 0
Frame #5: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #6: 0 0 1 0 0 0 0 0 0 0 0 0
Frame #7: 0 1 0 0 0 1 0 0 0 0 0 1
Frame #8: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #9: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #10: 0 0 0 0 0 0 0 0 0 0 0 0

Figure 3: 10 Frames each includes 10 bit sequence to encode "a"

Frame #1: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #2: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #3: 0 1 1 0 0 0 1 0 0 0 0 0
Frame #4: 0 0 0 0 0 0 0 0 0 0 1 0
Frame #5: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #6: 0 0 0 0 0 0 0 0 0 1 0 0
Frame #7: 0 1 0 0 0 0 1 0 0 0 0 0
Frame #8: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #9: 0 0 0 0 0 0 0 0 0 0 0 0
Frame #10: 0 0 0 0 0 0 0 0 0 0 0 0

Figure 4: 10 Frames each includes 10 bit sequence to encode "z"

After this step, I will perform frame clustering to specifying the similarities within different characters. For example, I will look at the frame #1 for each every character, and then choose those characters which have the same bit sequence in the frame. A summary of the similarities for the first and fifth frames are shown in Table 1.

Table 1: A summary of the similarities for the first and fifth frames

Frame Number	Similarities
#1	[0 0 0 0 0 0 0 0 0 0 0 0]: a, d, e, c, f, g, j, m, n, o, p, q, r, s, u, v, w, x, y, z [1 0 0 0 0 0 1 0 0 0 0 0]: b, h, k, t [0 0 0 0 1 0 0 0 0 0 0 0]: i [0 0 0 0 0 0 0 0 0 0 0 1]: e
#5	[0 0 0 0 0 0 0 0 0 0 0 0]: a, b, c, d, e, f, g, h, i, j, k, l, n, o, p, q, r, s, t, u, v, x, y, z [0 0 0 1 0 0 0 0 0 0 0 0]: m [1 0 0 0 0 0 1 0 0 0 0 0]: w

LVQ neural networks have a competitive layer a linear layer. The competitive layer learns to classify input vectors. The linear layer transforms the competitive layer's classes into target classification defined by the user. The classes learned by the competitive layer are referred to as subclasses and the classes of the linear layer as target classes. As I discussed in Section 2, I define 10 frames

in eachcharacter, specifying 12 bit sequence for every frame. I design ten LVQ networks. Each LVQ network has a 12*1 input vector. Based on the frame similarities, the output vector will produce for each LVQ network layer. Table 2 presents the output

Table 2: The output vectors for the first five layers. As you see, the dimension of output vectors depends on the values in Table 1

Frame Number	Output Vector
#1	4*1
#2	4*1
#3	12*1
#4	15*1
#5	3*1

vectors for the first five layers. As you see, the dimension of output vectors depends on the values in Table 1.

In general, I need twenty six networks to determine an English character. The reason is that there are 26 English alphabets which I like to detect.

Demo

A demo snapshot of the running program is presented in Figure 5.

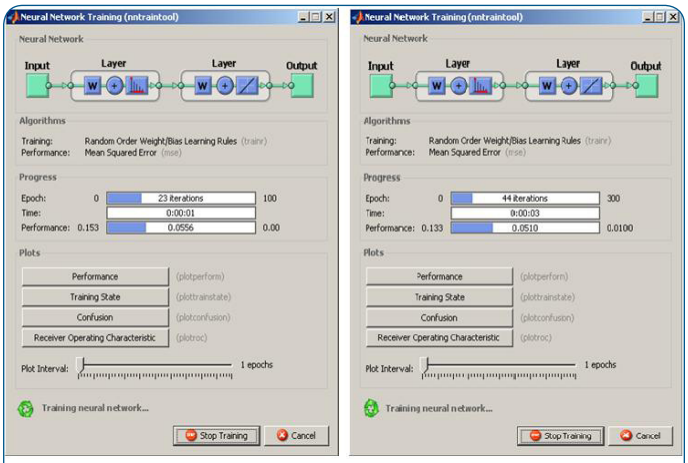

Figure 5: A sample run of the Matlab implementation

Experimental Validations

To prove the general performance of the proposed machine-written character detection method, some experiments using synthetic data were carried out. In particular, a system has been implemented to emulate the proposed approach. This system was built using Matlab R2013a. All the experiments were carried out on a 3.00 GHz Intel Dual core 2MB cache with 2GB of RAM running 64-bit Windows 7 operating system.

True and False Detection

In Table 3, you can see the percentage of true and false detection of each lower case character. In this experiment, I have performed the same implementationfor each alphabet, running for ten times. As you can see in this experiment, similar shapes may cause false detection. For instance, character "j" could be detected as "i" or vice versa, and character "c" might be recognized as "o". Different encoding strategy may recover existing false detection rates.

Table 3: True and false detection rate of the proposed system

Character	True Detection (%)	False Detection With
a	100	
b	100	
c	81	o
d	100	
e	93	c
f	100	
g	100	
h	100	
i	92	j
j	82	i
k	100	
l	100	
m	93	h
n	100	
o	87	c
p	100	
q	100	
r	100	
s	100	
t	100	
u	100	
v	100	
w	100	
x	100	
y	100	
z	100	

Time Consumption

In this experiment, I investigate the time consumption of the system regarding different single characters in the set (Table 4).

Discussion and Future Work

A supervised classification strategy based on the LVQ algorithm is presented to recognize English machine-written characters. In this contribution, I havedeveloped a particular encoding system combining with the LVQ algorithm to classify and characterize every lower case character, and then recognize a specific single character in a given set. The work demonstrates that the LVQ algorithm can be employed appropriately for machine-written character recognition. Two different experimental validations have also been done on synthetic data. These experimental results obtained from the method, specially the true and false alerts clearly shown the promising performance and reliability of the system.

As a future work, I plan to examine other supervised machine learning techniques along with advanced adaptive and evolutionary approaches [16-21] and investigate large and public data sets. Working on noisy data could be also another future direction.

Table 4: Time consumption of the proposed system regarding different single characters

Character	Time Consumption (Sec.)
a	4.268e+001
b	3.367e+001
c	2.114e+001
d	4.982e+001
e	5.093e+001
f	4.438e+001
g	5.125e+001
h	3.764e+001
i	2.946e+001
j	3.761e+001
k	4.836e+001
l	2.121e+001
m	5.337e+001
n	4.342e+001
o	4.504e+001
p	4.129e+001
q	4.306e+001
r	4.062e+001
s	4.651e+001
t	4.732e+001
u	3.361e+001
v	3.012e+001
w	4.963e+001
x	5.125e+001
y	4.502e+001
z	4.317e+001

References

1. Gersho A, Gray RM. Vector quantization and signal compression. Springer Science & Business Media. 2012; 159.
2. Baig MH, Rasool A,Bhatti MI. Classification of electrocardiogram using SOM, LVQ and beat detection methods in localization of cardiac arrhythmias. In Engineering in Medicine and Biology Society. 2001: Proceedings of the 23rd Annual International Conference of the IEEE; 2001; IEEE.Vol 2. p. 1684-1687.
3. Amezcua J, Melin P, Castillo O. Design of an Optimal Modular LVQ Network for Classification of Arrhythmias Based on a Variable Training-Test Datasets Strategy. In Intelligent Systems' 2014. Springer International Publishing. 2015: 369-375.
4. Mudali D, Biehl M, Leenders KL,Roerdink JB. LVQ and SVM Classification of FDG-PET Brain Data. In Advances in Self-Organizing Maps and Learning Vector Quantization. Springer International Publishing. 2016: 205-215.
5. Bekaddour A, Bessaid A,Bendimerad FT. Multi Spectral Satellite Image Ensembles Classification Combining k-means, LVQ and SVM Classification Techniques. Journal of the Indian Society of Remote Sensing. 2015;43(4): 671-686.
6. Ortiz A, Górriz JM, Ramírez J,Martínez-Murcia FJ.Alzheimer's Disease Neuroimaging Initiative. LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimer's disease. Pattern Recognition Letters. 2013;34(14): 1725-1733.
7. Mori M, Uchida S,Sakano H. Global feature for online character recognition. Pattern Recognition Letters. 2014; 35: 142-148.
8. Giménez A, Khoury I, Andrés-Ferrer J, Juan A. Handwriting word recognition using windowed Bernoulli HMMs. Pattern Recognition Letters. 2014; 35: 149-156.
9. Dash T. Time efficient approach to offline hand written character recognition using associative memory net. 2013; arXiv preprint arXiv: 1306.4592.
10. Mahajan J,Mahajan R. Designing an Intelligent System for Optical Handwritten Character Recognition using ANN. International Journal of Computer Applications. 2014; 91(13).
11. George HB, Alan SD.Int Computers Ltd 1971. Character recognition systems. US Patent 3,609,686.
12. Swiercz E. Automatic classification of LFM signals for radar emitter recognition using wavelet decomposition and LVQ classifier. Acta Phys. Pol. A. 2011; 119: 488-494.
13. Malakooti MV, Tafti AP,Naji HR. An efficient algorithm for human cell detection in electron microscope images based on cluster analysis and vector quantization techniques. In Digital Information and Communication Technology and it's Applications (DICTAP). 2012: Second International Conference; 2012: IEEE. p. 125-129.
14. Bardosi Z, Granata D, Lugos G, Tafti AP,Saxena S. Metacarpal Bones Localization in X-ray Imagery Using Particle Filter Segmentation. 2014; arXiv preprint arXiv: 1412.8197.
15. Melin P, Amezcua J, Valdez F, Castillo O. A new neural network model based on the LVQ algorithm for multi-class classification of arrhythmias. Information Sciences. 2014; 279: 483-497.
16. Shin Y, Lee S, Ahn M, Cho H, Jun SC, Lee HN. Simple adaptive sparse representation based classification schemes for EEG based brain–computer interface applications. Computers in biology and medicine. 2015; 66: 29-38.
17. Tongaonkar A, Torres R, Iliofotou M, Keralapura R,Nucci A. Towards self adaptive network traffic classification. Computer Communications. 2015; 56: 35-46.
18. Tafti AP, Kirkpatrick AB, Alavi Z, Owen HA, Yu Z. Recent advances in 3D SEM surface reconstruction. Micron. 2015; 78: 54-66.
19. Ruiz AB, Saborido R,Luque M. A preference-based evolutionary algorithm for multiobjective optimization: the weighting achievement scalarizing function genetic algorithm. Journal of Global Optimization. 2015; 62(1): 101-129.
20. Tafti AP, Kirkpatrick AB, Holz JD, Owen HA, Yu Z. 3DSEM: A 3D microscopy dataset. Data in brief. 2016; 6: 112-116.
21. Hu YY, Li DS. Bands selection and classification of hyperspectral images based on hybrid kernels SVM by evolutionary algorithm. In Selected Proceedings of the Chinese Society for Optical Engineering. International Society for Optics and Photonics; 2016.