A Novel Vector Quantization Approach to Arabic Character Recognition

Ahmad M. Sarhan, Ph.D. and Omar I. Al Helalat, M.Sc.

Abstract— **In this paper, a novel approach to Arabic letter recognition is proposed. The system is based on the classified vector quantization (CVQ) technique employing the minimum distance classifier. To prove the robustness of the CVQ system, its performance is compared to that of a standard artificial neural network (ANN)-based solution. In the CVQ system, each input letter is mapped to its class using the minimum Euclidean distance. Simulation results are provided and show that the CVQ system always produces a lower Mean Squared Error (MSE) and higher success rates than the current ANN solutions.**

Index Terms— **Backpropagation, word, codebook, ANN, neuron, Widrow-Hoff rule, MSE, Gaussian, standard deviation, classified vector quantization, minimum distance classifier**

I. Introduction

Statistical clustering methods have been widely used for a variety of signal processing applications, including classification and vector quantization (VQ). In this paper, a classified vector quantization (CVQ) scheme for the recognition of typed Arabic letters is developed. The CVQ technique used here employs the minimum distance classifier. Specifically, the minimum distance investigated here is the Euclidean or squared-error distance.

To prove the power of the proposed method, it is compared with the existing current solution to the problem, namely a backpropagation ANN system. Simulation results prove that the CVQ system always produces higher success rates than the standard ANN-based approach.

A. Overview

The goal of a letter recognition system is to transform a text document typed on paper into a digital format that can be manipulated by word processing software. The system is required to identify a given input data/letter form by mapping it to a single letter in a given letter set. This process can be quite involved since there are several valid forms that a letter may take. This is largely due to the many fonts and styles (bold type, italic type, etc.…) that can be used. The motivation behind developing letter recognition systems is inspired by their wide range of applications including archiving documents, automatic reading of checks, and number plate reading.

B. Arabic letter recognition

Arabic belongs to the group of Semitic alphabetical scripts in which mainly the consonants are represented in writing, while the markings of vowels (using diacritics) is optional and is rarely used. Arabic is spoken by more than 300 million people and is the official language of many countries.

Enormous amount of research has been undertaken in the field of recognizing typed and handwritten Latin, Chinese, and Indian letters. Little progress, however, has been made in the recognition of Arabic letters, mainly due to their cursive nature. Unlike most of the other languages, both typed and hand-written Arabic letters are cursive. Furthermore, Arabic letters can take more shapes than Latin letters.

Other problems facing Arabic letter recognition systems include:

a. The unevenness of Arabic fonts; i.e., a certain letter in a specific font can be misinterpreted as a different letter in another font. In Arabic, some letter pairs may be combined together to form another letter, that is often referred to as a ligature. The only mandatory ligature is the (Lam Alef). Other ligatures are optional. Ligatures greatly complicate the segmentation task of an Optical Character Recognition (OCR) system.

b. Arabic has 28 letters, each of which can be linked in three different ways or separated depending on the case. Therefore, each letter can have up to four different forms depending on its position.

c. Arabic letters have different heights, which puts an extra burden on the noise detection task of the OCR system.

d. Line mingling, a phenomenon exhibited by improperly spaced documents.

ISBN:978-988-98671-5-7

II. The Vector Quantization Partitioning (VQP) Scheme

Introduced and discussed in this section is the *VQP* scheme, a partitioning method that partitions the space R^M into several partitions based on *VQ*. In our character recognition application, the partitions are used to define the classes.

A. Overview

Statistical clustering methods have found many applications in signal processing including both classification and VQ. In recent years, VQ has become a very popular technique for many applications involving data compression. This is due to the fact VQ produces a lower distortion than scalar quantization for a given rate. Another property of VQ that makes it especially useful as a partitioning scheme is its preservation of correlation between vector components; and hence, the ability to remove spatial or temporal redundancies between blocks of samples.

B. Vector Quantization (VQ)

A z-level vector quantizer **Q** is a mapping of each input vector $\mathbf{y} \in R^M$ to a vector $\mathbf{z} \in R^M$ called a codeword, or a reproduction vector, drawn from a finite point-set, $C = \{\mathbf{z}_i, i=1,\ldots,z\}$, called a codebook. A unique index i=1,2,…, z is associated with each codeword. Clearly, the codewords, z_i, i=1,2,…,z, partition the space R^M into z regions s_i, i = 1,2,…,z such that

$$\mathbf{s_i} = \{ \mathbf{y}: \mathbf{Q}(y)=i\} \qquad (1)$$

In our application, a codeword refers to an Arabic letter, and a codebook refers to the set of all Arabic letters.

To encode an input vector, or a feature vector representing an Arabic letter, the quantizer measures the distortion between the input vector and every reproduction vector from the codebook. The reproduction vector yielding the minimum distortion is used as the output vector (class).

Many distortion measures have been proposed in the literature including the squared-error distortion; the l_v or Holder norm; and l_∞ norm. The most commonly used distortion measure, however, is the squared-error, or *Euclidean*, distance. Here, the partitions are called the Voronoi regions and are given by

$$s_i = \{y: \| y-z_i \|^2 \leq \| y-z_i \|^2, j = 1,2,\ldots,z, j \neq i\} \qquad (2)$$

A simple way to generate the codebook is by using the LBG algorithm [18].The LBG algorithm, proposed by Linde et al, is commonly used in the design of an optimal codebook from empirical data. In our character recognition application, a good choice for the initial codewords required by the LBG algorithm is the vectors representing the 28 Arabic letters.

The Euclidean distance between two n-dimensional (row or column) vectors **x** and **y** is defined as the scalar

$$d(\mathbf{x}, \mathbf{y}) = \| \mathbf{x}-\mathbf{y} \| = \| \mathbf{y}-\mathbf{x} \| = [(x_1-y_1)^2+\ldots+(x_n-y_n)^2]^{1/2} \qquad (3)$$

This expression is simply the norm of the difference between the two vectors

It is necessary to compute a set of Euclidean distances between vector **y** and each member of a vector population consisting of p, n-dimensional vectors arranged as the rows of a p x n. For the dimensions to line up properly, **y** has to be of dimension 1 x n. Then the distance between **y** and each element of **x** is contained in the p x 1 vector

It is not difficult to show that selecting the smallest distance is equivalent to evaluating the functions

$$d_j(\mathbf{x}) = \mathbf{x}^T \mathbf{m}_j - (1/2)\, \mathbf{m}_j^T \mathbf{m}_j \qquad j = 1, 2,\ldots, C, \qquad (4)$$

and assigning **x** to class c_i if $d_i(\mathbf{x})$ yields the largest numerical value.

The decision boundary between classes C_i and C_j for a minimum distance classifier is

$$d_{ij}(\mathbf{x}) = d_i(\mathbf{x})-d_j(\mathbf{x}) = \mathbf{x}^T(\mathbf{m}_i-\mathbf{m}_j) - 0.5\ (\mathbf{m}_i-\mathbf{m}_j)^T(\mathbf{m}_i+\mathbf{m}_j) = 0 \qquad (5)$$

The surface given by this equation is the perpendicular bisector of the line segment joining $\mathbf{m}_i$ and $\mathbf{m}_j$. For n = 2, the perpendicular bisector is a line. For n = 3 it is a plane, and for n>3 it is called hyperplane.

III. Artificial Neural Networks

ANNs were introduced by McCulloch and Pitts in 1943. ANNs are trainable algorithms that can "learn" to solve complex problems from training data that consists of a set of pairs of inputs and desired outputs (targets). They can be trained to perform a specific task such as prediction, and classification. ANNs have been applied successfully in many fields including speech recognition, image processing, and adaptive control.

An ANN consists of interconnected processing elements called *neurons* that work together to produce an output.

A. Single Neuron and the Least-Mean-Square (LMS) Algorithm

The neuron (Fig. 1) is the basic building block of an ANN.

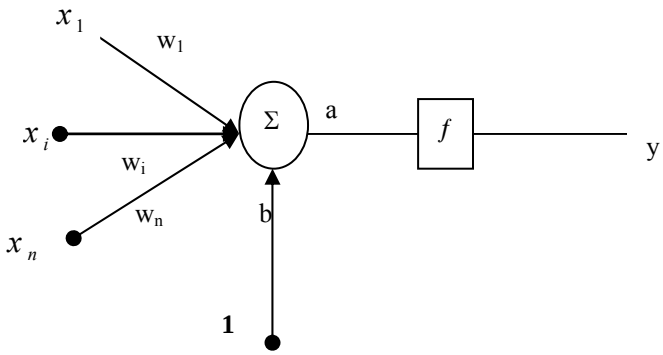
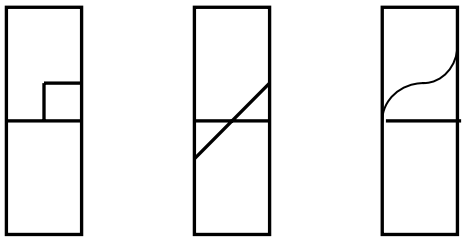


Figure 1*: The Structure of a single neuron*

The output *a,* of the neuron is a weighted linear combination of its inputs. The function *f* in Fig.1 is called the transfer or scaling function. Some commonly used transfer functions are shown in Fig.2.



(a) (b) (c)

Figure 2: Transfer *functions: (a) hard line (b) pure line, and (c) logsigmoid*

Using training data (input—target pairs), the weights of the neuron can be iteratively adjusted to give local or global optima. Optimum weights in the sense of Least Squared Errors were derived by Widrow and Hoff [9] and the algorithm was called the LMS algorithm and is commonly known as the *Widrow-Hoff rule* and has become a widely accepted algorithm .In the LMS algorithm, the network weights are moved along the negative of the gradient of the performance function.

Specifically, after each iteration or epoch (new set of input—target pairs) the weights are adjusted according to the following rule

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \mu\, e\, \boldsymbol{x}, \qquad (6)$$

where μ is the learning /adaptation speed, and the input vector $\mathbf{x} \in \mathbf{R}^n$ is given by

$$\boldsymbol{x} = [x_1, x_2, x_3, x_4, x_5, \ldots, x_n]^T, \qquad (7)$$

and $\mathbf{w} \in \mathbf{R}^n$ is the vector of weights and is given by

$$\boldsymbol{w} = [w_1, w_2, w_3, w_4, \ldots, w_n]^T. \qquad (8)$$

The output *a* of the neuron is defined by the following expression

$$a = w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_n x_n + b, \qquad (9)$$

where *b* is a bias value that is not part of the input. The neuron's scaled output *y* is given by $y = f(a)$, where *f* is the transfer function (Fig. 2). The error *e* is the difference between the neuron's output and the desired output.

The initial values of the weights can be set explicitly if apriori information is available. Alternatively and in most practical cases, the weights are initially set to zeros or some random values.

B.Multilayer ANN

In general, a multilayer ANN has the architecture depicted in Fig. 3, which shows a two-layer network

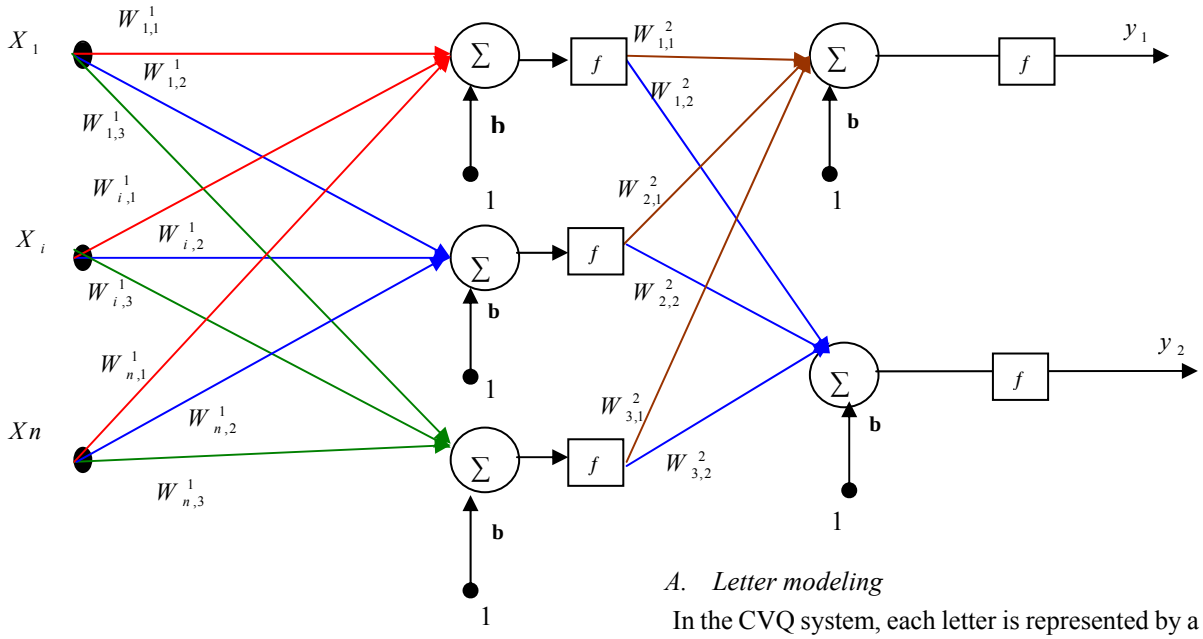


Figure 3: *A two-layer ANN*

ANN is designed so that the application of a set of inputs produces a desired set of outputs. One of the main letteristics of an ANN is its set of weights.

Note that $w^{k}_{i,j}$ in Fig. 3denotes the weight for the i^{th} input in the j^{th} neuron of the k^{th} layer.

C. Backpropagation Algorithm

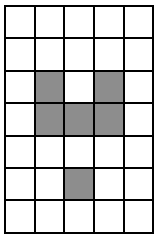
When a multilayer ANN uses the Widrow-Hoff learning rule and nonlinear differentiable transfer functions, the resulting ANN, known as Backpropagation, can approximate almost any function that has a finite number of discontinuities. Properly trained backpropagation networks have proven to give reasonable answers when presented with inputs that they have never seen. In many cases, it is possible to train a network on a representative set of (input—target) pairs and get good results without training the network on all possible (input—target) pairs.

IV. Methods

In the CVQ system (Fig.5), the set of all letters constitute the codebook. Arabic has 28 letters constituting 28 classes or 28-element codebook. Each letter is represented by a 35 element binary vector. For each input letter, the Euclidean Eq. 2 distance between that input vector representation and each vector in the codebook is calculated. (Eq.2).The index/class of the vector that corresponds to the minimum distance is used as the system's output. This output/classifier is commonly called the CVQ.

A. Letter modeling

In the CVQ system, each letter is represented by a matrix of 7 x 5 binary pixels, producing a 35-element input vector. As an example, Fig. 4 shows the binary image of the Arabic letter "Baa" and its corresponding matrix representation. The number of Arabic letters is 28, producing 28 different classes or system outputs.



$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(a) (b)

Figure 4: *The Arabic letter "Baa" (a) Bit map image, and (b) matrix representation*

The flow of the system is as follows. The input to the feature extraction stage consists of a vector of 35 elements representing an Arabic letter. The output of the feature extraction stage is a 36-element vector, with the additional element being the standard deviation of the original 35 elements. The 36-element vector is the fed as input to the ANN stage.

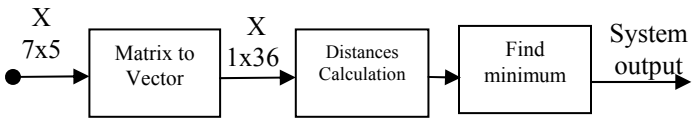


Figure 5: CVQ Flow diagram

Figure 6: Flow diagram of a standard ANN network

B.ANN Design

It was shown in [1] and [17] that the best ANN solution for the problem at hand is a two-layer network (1 hidden and 1 output) with 10 neurons in the hidden layer. The same network is also used here in the computer simulations to provide performance comparison with the CVQ system. Specifically, the ANN used is a Backpropagation network, i.e., it uses the LMS algorithm (Widrow-Hoff rule). Furthermore, log-sigmoid functions, Eq. 8, were used as the transfer functions of the output layer since they can approximate binary values.

$$\log sig(n) = \frac{1}{1+e^{-n}} \qquad (8)$$

The network receives the 36 Boolean values as a 36-element input vector. It has 28 outputs, where each output corresponds to an Arabic letter.

V. Simulations

The CVQ system was tested with sets of noisy and clean inputs. The contaminating noise was the Gaussian noise which has a probability density function (pdf) given by

$$f(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \qquad (9)$$

where μ and σ are the mean and standard deviation of the noise, respectively.

Fig. 7 shows the percentage of success rates as a function of noise level for the CVQ system. Clearly, the CVQ system (Fig.8) is able to produce high success rates even for highly contaminated inputs.

To further investigate the performance of the system and to provide some sort of merit, it is compared to that of an ANN . The network was a 2-layer backpropagation, with sigmoid transfer functions in the output layer as demonstrated in the literature [9].

Both systems were presented with clean and noisy inputs, where the noise was additive Gaussian noise of zero mean and varying standard deviation (noise level),. Fig.8 shows the performance of both systems for varying noise level. As can be seen from Fig.8, the CVQ always outperforms the standard ANN system and produces higher success rates. This is true for levels of contaminating noise.

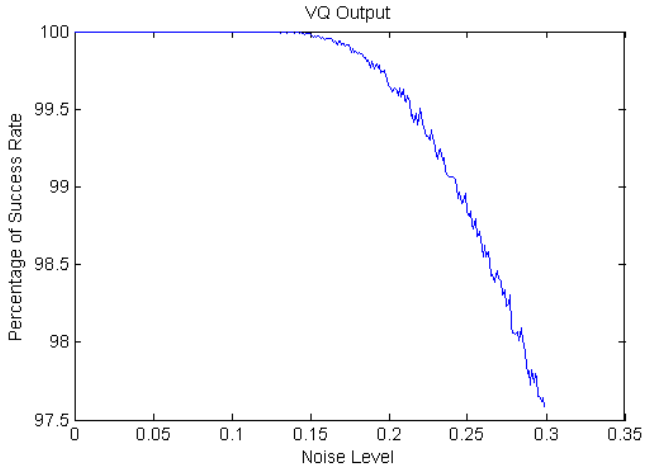


Figure 7: Percentage of success rate vs. Noise level for the CVQ system

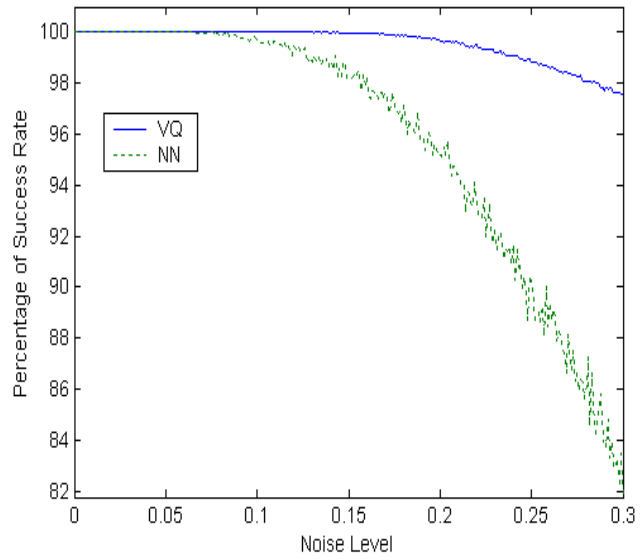


Figure 8: Percentage of success rate vs. noise level for the CVQ and the ANN systems

VI. Conclusion

In this paper, a novel approach to the recognition of typed Arabic letters is presented. The system is based on the classified vector quantization technique, employing the minimum distance classifier. The minimum distance studied here is the Euclidean distance.

The system inputs consist of the typed Arabic letters, where each letter is represented by a matrix of 7 x 5 binary pixels. Arabic has 28 letters constituting the range of system‘s outputs.

The CVQ system is compared to the standard solution that uses a backpropagation Artificial Neural Networks for classification. In addition to being able to always produce higher success rates, the CVQ system has many other advantages over the ANN system. For example, unlike an ANN system, the CVQ system does not require any training. A consequence drawn from this fact is that the CVQ system performs equally well on all inputs and is not biased towards a subset of the input (training data in the ANN case). ANN, just as all systems that require training, do not usually exhibit the same performance on data that they have never seen before.

Simulation results indicate clearly that the CVQ system always produces a lower MSE and higher success rates than the current ANN solutions for all levels of contaminating Gaussian noise.

References

[1] Amin and G. Masini, "Machine recognition of multifont printed Arabic texts", in Proc.8th Int. Conf. Patt. Recogn. (Paris, France),pp. 392-395, 1986

[2] A.M. Sarhan and R. C. Hardie, "Partition-based filters", In *Proceedings of the 1995 IEEE National Aerospace and Electronic Conference (NAECON),* volume 2, Dayton, Ohio, May 1995.

[3] A.M. Sarhan, R. C. Hardie, and K. E. Barner, "Partition-based adaptive estimation of single-response evoked potentials", In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* May 1995.

[4] A.M. Sarhan. N*onlinear partition-based filters for signal restoration.* Ph.D. Thesis. University of Dayton, July 1996.

[5] D. J. Hand. *Discrimination and Classification*. New York: Wiley, 1981E. W. Brown, "Letter Recognition by Feature Point Extraction", Northeastern University internal paper, 1992.

[6] F. Hussain and J. Cowell, “Letter Recognition of Arabic and Latin Scripts”, *Proceedings, IEEE International Conference on Information Visualization* ,pp. 51–56, 2000

[7] H. Al-Yousefi and S. S. Udpa, "Recognition of handwritten Arabic letters," in Proc. SPIE 32nd Ann. Int. Tech. Symp. Opt. Optoelectric Applied Sci. Eng. (San Diego, CA), Aug. 1988.

[8] Haykin S.,Adaptive Filter Theory, Englewood Cli s.N.J: Prentice Hall(3ed) ,1996

[9] J.F Canny. “A Computational Approach to Edge Detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, pp. 679-698, 1986.

[10] K. E. Barner, A. M. Sarhan, and R. C. Hardie. "Partition-based weighted sum filters for image restoration". *IEEE Transactions on Image Processing,* Vol. 8, No. 5, May 1999.

[11] K. Khatatneh,"Probabilistic Artificial Neural Network for Recognizing the Arabic. Hand Written Letters", Journal of Computer Science 3 (12), 881-886, 2006

[12] K. Badi and M. Shimura, "Machine recognition of Arabic cursive script" Trans. Inst. Electron. Commun. Eng., Vol. E65, no. 2, pp. 107-114, Feb. 1982.

[13] K. Badi and M. Shimura, "Machine recognition of Arabic cursive scripts" in Pattern Recognition in Practice. Amsterdam: North Holland, 1980.

[14] L.Hammami and D.Berkani ,"Recognition system for printed multi-font and multi-size Arabic letters", The Arabian Journal for Science and Engineering, Volume 27, Number 1B, pp:57-72, April, 2002

[15] M Altuwaijri, M.A Bayoumi, "Arabic Text Recognition Using Neural Network" ISCAS 94. IEEE International Symposium on Circuits and systems, Volume 6, 30 May-2 June 1994.

[16] N. Ben Amor, N. Essoukri Ben Amara: "A hybrid approach for Multifont Arabic Letters Recognition", 5th WSEAS Int.Conf. On Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'06) Madrid, Spain, February 15-17, 2006.

[17] R. A. Dosari, R. C. Hardie, and A. M. Sarhan. "Multi-channel nonlinear filters for signal restoration". In *Proceedings of the 1997 IEEE National Aerospace and Electronic Conference (NAECON),* Dayton, Ohio, May 1997.

[18] Y.Linde, A.Buzo, and R.M.GrAY.An algorithm for vector quantization. IEEE Transactions on Communication Theory, 28(1):84-95, January 1980.