Proceedings of the International Conference on Electrical and Computer Systems
Ottawa, Ontario, Canada, 22-24 August 2012
Paper No. 134

ANN's vs. SVM's for Image Classification

Poobalan Govender
Optimisation Studies Unit, Dept. of Electronic Engineering, Durban University of Technology
Steve Biko Road, Berea, Durban, South Africa
poobieg@dut.ac.za

[1]Neelendran Pillay, [2]Kevin Emmanuel Moorgas
Optimisation Studies Unit, Dept. of Electronic Engineering, Durban University of Technology
Steve Biko Road, Berea, Durban, South Africa
[1]trevorpi@dut.ac.za; [2]kevinm@dut.ac.za

Abstract –In this paper the dynamic performance of the artificial neural network is compared to the performance of a statistical method such as the support vector machine. This comparison is made with respect to an image classification application where the performance is compared with regards to generalization and robustness. Image vectors are compressed in order to reduce the dimensionality and the salient feature vectors are extracted with the principle component algorithm. The artificial neural network and the support vector machine are trained to classify images with feature vectors. A comparative analysis is made between the artificial neural network and the support vector machine with respect to robustness and generalization.

Keywords**:** support vector machine, artificial neural network, principle component analysis, hyperplane

1. Introduction

Texture and tone are two salient aspects of an image and on occasion the one property can easily overlook the other (Seetha et al., 2005-2008). Texture depends on the spatial distribution of gray levels within a neighborhood (Tomasi and Manduchi, 1998). Texture displays its characteristics by means of pixels, and pixel values and images can be identified by their texture (Chen et al., 1993). Artificial neural networks (ANN's) and support vector machines (SVM's) have been widely used for image identification (Seetha et al., 2005-2008). Image identification using ANN's and SVM is achieved by extracting key textural characteristics during image pre-processing using techniques such as wavelet compression (WC) and principle component analysis (PCA), and then training the classifier to identify these characteristics. In this paper we will focus on the design and performance of ANN's and SVM's for performing reliable and repeatable image identification under varying degrees of operational challenges.

2. The ANN

The ANN has been chosen because of its ability to provide solutions to problems that are characterized by high dimensionality noise, nonlinearities and error prone data (Haykin, 1999). The artificial neuron in Fig. 1 is the elementary processing unit of an ANN. Each connection to a neuron is defined by a weight (w). An activation function (f) shapes the output of the neuron (y) before it is applied to the next neuron, or forms the output. Neurons combine to form either feedforward (FF-ANN) or recurrent networks, and variants thereof.

Our study uses a feed-forward ANN architecture (FF-ANN) that has been designed and trained for image recognition. The tan-sigmoid output transfer function was chosen because of its universal applicability to a wide range of problems (Kilian, 1996). FF-ANN's consist of three or more layers, namely an input layer, one or more hidden layers, and an output layer (Haykin, 1999). Signal propagation is unidirectional from input to output. Image identification with ANN's is done by training the ANN to

recognize certain key textural features of an image. Textural features are detected using both pixels and pixel values, and is characterized by the spatial distribution of gray levels within a neighbourhood.

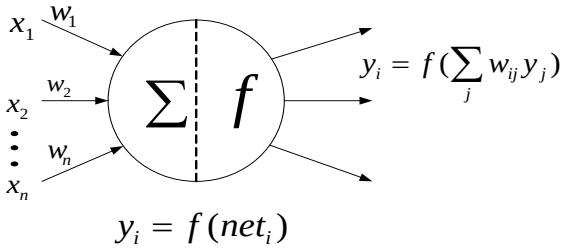


Fig. 1. **A**rtificial neuron basic processing element

3. Basic Theory of the SVM

SVM is a statistic based pattern classification technique introduced by Vapnik (1995). SVM is based on the concept of structural risk minimization (SRM). A learning machine's risk (*R*) is bound within the sum of the empirical risk (*Remp*) and a confidence interval ψ i.e. $R \leq Remp + \psi$ (Vapnik, 1995). SVM's utilize a kernel function to map a nonlinearly separable vector into a higher dimensional space to make it linearly separable. The kernel function is one of the main building blocks of a SVM and can be used with a wide range of different learning theories. When we draw a comparison between ANN's and SVM's, the choice of a suitable ANN architecture for a specific application is analogous to choosing a suitable kernel function for a SVM. Single SVM's are binary classifiers that can be extended by integrating several together for solving multiclass data problems (David and Lerner, 2004).

3.1 The optimal Separating Hyperplane

The SVM algorithm is used to determine the optimal separating hyperplane between two classes of data. Assume dataset *T* has two separable classes and a total of *k* samples, where these samples are represented as (x_1, y_1), (x_2, y_2),…, (x_k, y_k). The class label is represented as $y \in \{-1,1\}$ and is the binary value of two classes; $x \in R^n$ where R^n is an *n*- dimensional space.

Consider the SVM system in Fig. 2: Dataset *T* is separated into two classes by two parallel hyperplanes H_1 and H_2. Hyperplane H_3 is the optimal separating hyperplane that lies parallel with H_1 and H_2 and is equidistant between these two hyperplanes. H_3 is defined as $w \cdot x + b = 0$, where $\cdot$ denotes matrice multiplication, *w* is the normal vector to H_1 and H_2 , and *b* represents the bias.

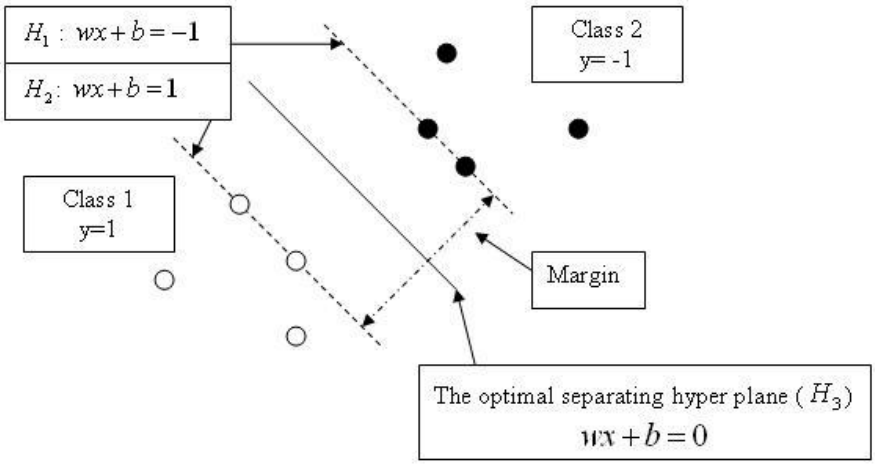


Fig. 2. Optimal separating hyperplane for a 2-D and 2-class problem

3.2 The Kernel Function

The kernel function is the key component of the SVM classifier and locates the decision boundaries between different classes of data (Cristianini and Shawe-Tayler, 2000). It converts training data from a feature space to a higher dimension and makes them linearly separable. Commonly used kernel functions include the polynomial function, the radial basis function and the sigmoid function. Binary class data being classified must meet the following condition:

$$\left.\begin{array}{l} wx_i + b \geq 1, 1 \leq i \leq k \text{ and } \forall_i = 1 \\ wx_i + b \leq -1, 1 \leq i \leq k \text{ and } \forall_i = -1 \end{array}\right\} \Rightarrow y_i(wxi + b) \geq 1, \forall_i = 1, 2, ..., k \quad (1)$$

The margin m in Fig. 2 represents the absolute distance between hyperplanes H_1 and H_2 and is denoted as $m = \frac{2}{\|w\|}$ (Vapnik,1998). This sets the restriction condition of the optimal separating hyperplane for ensuring that m obtains its maximum value. If we subject $m = \frac{2}{\|w\|}$ to (1) and replace the maximization of $m = \frac{2}{\|w\|}$ with its equivalent minimization of $\frac{\|w\|^2}{2}$, which is then solved by the following Lagrange formulation:

$$L(w,b,a) = \frac{\|w\|^2}{2} - \sum_{i=1}^{k} a_i [y_i(wx_i + b) - 1], a_i \geq 0, 1 \leq i \leq k \quad (2)$$

where a_i is the Lagrangian multiplier. To minimize $L(w,b,a)$ we minimize w and b as follows:

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{k} a_i x_i y_i \text{ and } \frac{\partial}{\partial b} L = 0 \Rightarrow b = \sum_{i=1}^{k} a_i y_i = 0 \quad (3)$$

and a_i is maximized as follows:

$$L(a) = \sum_{i=1}^{k} a_i - \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j y_i y_j (x_i x_j) \quad (4)$$

for any $i = 1, ..., n$. The kernel (k) is defined by $k(x_i,x_j) = x_i \cdot x_j$.

The quadratic is subjected to

$$\sum_{i=1}^{k} a_i y_i = 0 \text{ where } a_i \geq 0 \, \forall_i = 1, ..., k \quad (5)$$

The construction of the optimal separating hyperplane depends on solving the quadratic programming problem with equation (4) and equation (5). For the solution just a few of a is non-zero and its corresponding samples H_1 and H_2 in Fig. 2 form the support vectors. The SVM classifier is concluded as

$$f(x) = sgn(w \cdot x + b = sgn\left(\sum_{\forall_i a > 0} y_i a_i (x_i x) + b\right)) \quad (6)$$

For many real classification problems, the data vector set is usually nonlinear separable in the low dimension. For such instances the SVM classifier must be trained to solve the classification problem.

4. Image Pre-processing

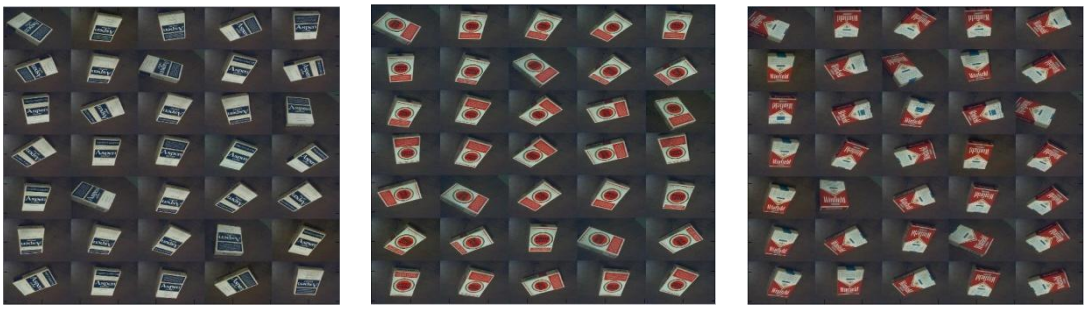
Image pre-processing involves image capture, image compression and feature extraction (Li et al., 2008)

4.1 Image Capture

The cigarette carton images considered in the study are given in Fig. 3. The identification of these images will be performed first by using the ANN and then the SVM. A comparison will then be made between the SVM and the ANN with regards to their ability to generalize and remain robust in the presence of environmental disturbances.

4.2 Data Compression with the Haar Wavelet Transform

Our image recognition system uses the Haar wavelet transform (HWT) for data compression to reduce the dimensionality of image data. Haar wavelet compression was chosen over other traditional compression methods because of the transform's ability to provide a multi-resolution representation of an image and to yield a higher compression ratio (Raviraj and Sanavullah, 2007), (Seetha et al., 2005-2008). In wavelet compression, an input signal is decomposed into a summation of a series of base functions or wavelets that are generated through dilating and shifting operations from a mother wavelet (Zhang et al., 2000).



(a) Aspen (b) Lucky Strike (c) Winfield

Fig. 3. Box Images in different positions

4.3 Feature Extraction Using PCA

PCA is a classical statistical data reduction technique that reduces the dimensionality of dataset vectors by identifying and extracting only key features of the dataset (Nixon and Aguado, 2008). Reduced datasets ease the computation burden during data processing (Nixon and Aguado, 2008). Following 2D-HWT compression, PCA extracts the most significant elements of the compressed image data set as follows:

Determine the mean of the compressed data set and subtract its mean from each data element to get the average. This produces a data set having a mean of zero and

i) Determine the covariance matrix, and then calculate the eigenvectors and eigenvalues of the covariant matrix.

Each column of the eigenvector matrix with the highest eigenvalues are the principal components and form the feature vector set.

5. Extracting Training Vectors and Ann and Svm Design

5.1 Extracting the Training Vectors

A comprehensive set of image vectors for each workpiece is determined as follows:

i) 3 images of each object from 120^0, 240^0 and 360^0 orientations are captured of each object when the object is in a first *pose*. The 3 images of the object in its first pose are clustered together to form a group.

ii) The pose of each object is adjusted 5 times and 3 images are taken of the object in each respective pose. From this we will have 5 groups of images, with each group having three images of the same object positioned in a certain pose.
iii) Repeat (i) and (ii) for each object. This results in a total of 15 groups of vectors (5 for each object), with each group having three images of each object occupying a specific pose.
These object vectors have different position orientations and poses to ensure that accurate recognition will always take place even under dynamic environmental conditions. Following HWT compression of the object image in each of its poses, PCA is applied to the compressed image in order to extract the salient feature vectors given in Table 1. These vectors are used to train the ANN and the SVM system.

5.2 Design of the Ann Recognition System

The ANN system used to perform the recognition function is given in Fig. 4. The size of the input layer is determined from the number of rows in the eigenvector data matrix, which is 36 from Table 1. The decision to use 9 neurons in the hidden layer was made following a series of experiments. The output layer consists of 3 neurons to correspond to the number of objects that must be identified for sorting.

5.3. Construction of the SVM Classifier

SVM's use optimization algorithms to determine the optimal boundaries between different classes of data (Cristianini and Shawe-Tayler, 2000; Chen et al., 1993) . The three cigarette cartons in Fig. 1 were used to generate the three classes of data used in the study. The SVM classifier can only classify two different data samples at a time. For this reason we converted our multi-class classification problem into two binary class problems and designed a multi-level classifier with two SVM classifiers for classifying the images of the three different cigarette brands (See Fig. 5).

5.3.1 SVM Tree Structure

Fig. 5 illustrates the SVM decision engine used in this study. From Fig. 5, the test sample on Level 1 contains three data classes that correspond to the Aspen, Lucky-Strike and Winfield cartons respectively. The SVM is a binary classifier and can only differentiate between two classes of data at a time. For this reason at Level 1 the SVM is trained to recognize the Aspen-Lucky-Strike vector combination as the one group and the Winfield carton vectors as the other group. At Level 2 the SVM is trained to differentiate the Lucky-Strike from the Aspen carton.

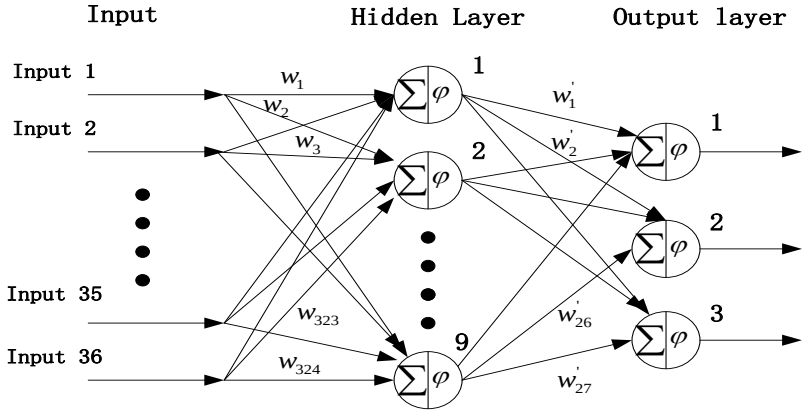


Fig. 4. 36: 9: 3 MLFF sigmoidal ANN for the cigarette carton recognition system

Table 1. *36 x 15* feature vectors to train the ANN and the SVM

	Aspen					Lucky Strike					Winfield				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-2.5677	-2.777	-2.777	-2.5511	-2.9558	-2.9914	-3.2594	-3.0475	-3.4609	-3.0522	-2.9195	-3.2473	-2.8643	-3.1482	-3.1322
2	1.3008	1.9072	1.9664	1.9593	1.7795	1.1517	1.2529	1.2451	1.3712	1.2356	1.1566	1.2594	1.1468	1.1946	1.2093
3	0.7221	0.4633	0.273	0.3233	0.6372	0.9604	1.024	0.9175	1.0769	0.9324	1.0105	1.071	0.9763	1.0627	1.0359
4	0.5448	0.4065	0.232	0.2685	0.5391	0.8794	0.9825	0.8849	1.0127	0.8842	0.7524	0.9168	0.7413	0.8908	0.8871
5	-0.3162	-0.7643	-0.9216	-0.8835	-0.5827	-0.0699	-0.0814	-0.1229	-0.1063	-0.114	0.0691	-0.0773	-0.0801	-0.0551	-0.0706
6	-1.6199	-1.6743	-1.4694	-1.5198	-1.753	-1.2539	-1.4084	-1.419	-1.4986	-1.3963	1.4789	-1.4594	-1.452	-1.3961	-1.3906
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
31	1.2972	0.5029	1.1143	1.1377	0.382	0.6042	0.4233	0.5308	0.499	0.5657	0.9309	0.7629	0.8867	0.8246	0.7929
32	0.7247	0.4491	0.7200	0.7479	0.2833	0.4376	0.177	0.4785	0.1256	0.4527	0.5874	0.0976	0.6075	0.742	0.3304
33	-0.1247	-0.0012	-0.0432	0.7200	-0.0012	-0.0048	-0.0021	-0.0014	-0.0036	-0.0031	-0.0314	-0.0145	-0.0259	-0.0059	-0.0206
34	-0.1533	-0.0135	-0.1212	-0.122	-0.0354	-0.0567	-0.0877	-0.0197	-0.1429	-0.0421	-0.1121	-0.2553	-0.0909	-0.0233	-0.1691
35	-0.8536	-0.3543	-0.8127	-0.8548	-0.3408	-0.5068	-0.2767	-0.5631	-0.2876	-0.5521	-0.7033	-0.3672	-0.7034	-0.673	-0.533
36	1.1316	0.369	0.977	1.0205	0.3773	0.5683	0.3666	0.5841	0.4335	0.5974	0.8468	0.637	0.8202	0.7023	0.7227

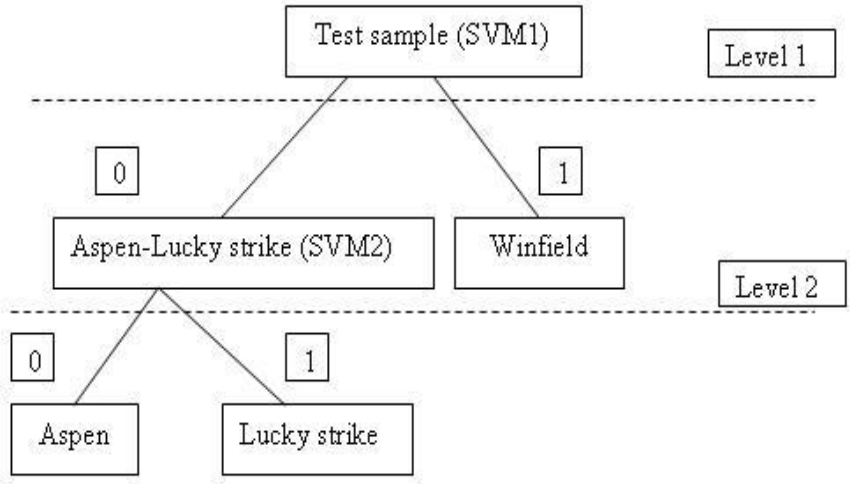


Fig. 5. Structure of the multiclass SVM decision tree

5.3.2 Selection of the SVM Kernel Function

The performance of a SVM depends largely on its kernel function and proper selection of the kernel function will determine how well the SVM generalizes (Cristianini and Shawe-Tayler, 2000). However, there is no set theoretical technique to determine a kernel function. For this reason we followed an iterative process to try the different kernel functions, and finally chose a linear kernel function based on its performance for the Level 1 and Level 2 SVM classifiers.

5.3.3 SVM Classification Results

Table 2 shows the results of the SVM1 classifier. The 0 denotes the target vectors of the Aspen-Lucky-Strike combination and 1 denotes the Winfield carton target vectors. On Level 2, the SVM2 classifier differentiates between the Aspen and Lucky Strike cartons. In Table 3 for the SVM2 classifier, the 0 denotes the Aspen carton and the 1 represents the Lucky Strike carton. From these results we can conclude that the three different boxes were recognized successfully with a recognition rate of 100%.

6. Comparing the Dynamic Performance of the ANN to that of the SVM

6.1 Generalization Ability

The SVM and NN were first compared with regards to their ability to generalize. Image data was used as training data for the NN and SVM classifier. We used five ANN classifiers and five SVM classifiers to test for generalization. These were trained using 3, 6, 9, 12 and 15 samples respectively for

each cigarette boxes, and each test sample had 20 images of each cigarette box. From Fig. 6, for small quantities of training set samples the SVM classifier has stronger generalization ability than the ANN classifier for the same training conditions. For larger training data sets, the performance of the SVM and NN are comparable and reinforces the fact that ANN's usually produce good results when large quantities of data are available.

6.2 Robustness

A comparison between the SVM and the ANN classifier was also done to test their ability to reject noise disturbances from external environmental factors. Artificial 'salt and pepper' noise was used to test the immunity of the SVM and ANN classifiers to noise interferences. Varying degrees of 'salt and pepper' noise was applied to 15 test sample images to compare the robustness of the SVM classifier to that of the ANN. The results of these tests are given in Fig. 7 and Fig. 8. From Fig. 7 we observe the following: For small quantities of data and low noise levels (See Fig. 7) the SVM exhibits slightly better robustness than the ANN. When large data quantities are available the ANN outperforms the SVM.

Table 2. Level 1 classification for SVM1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Table 3. Level 2 classification for SVM2

1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	1	1	1	1	1

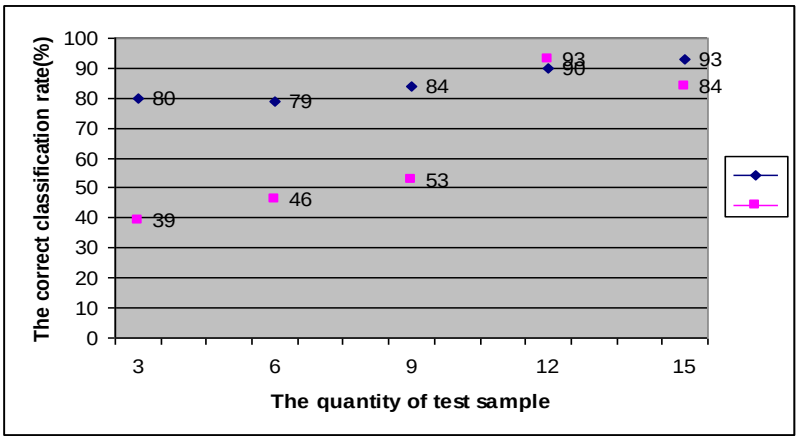


Fig. 6. Generalization performances of the ANN and SVM classifiers
(The diamond represents the SVM classifier; the square represents the ANN classifier)

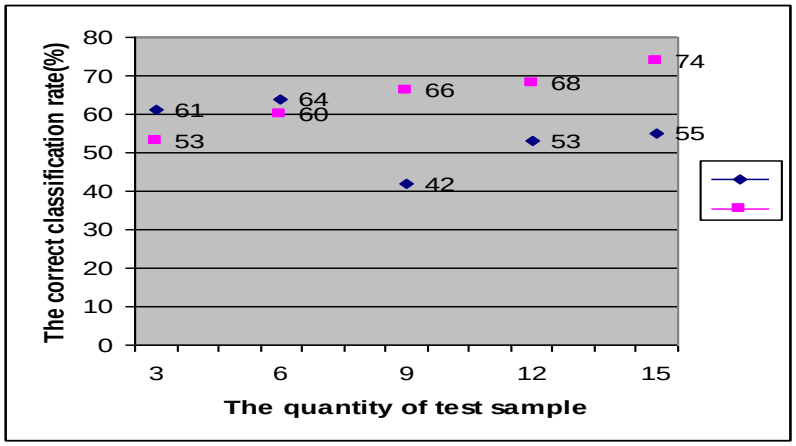


Fig.7. Robustness test with level 0.2 salt and pepper noise

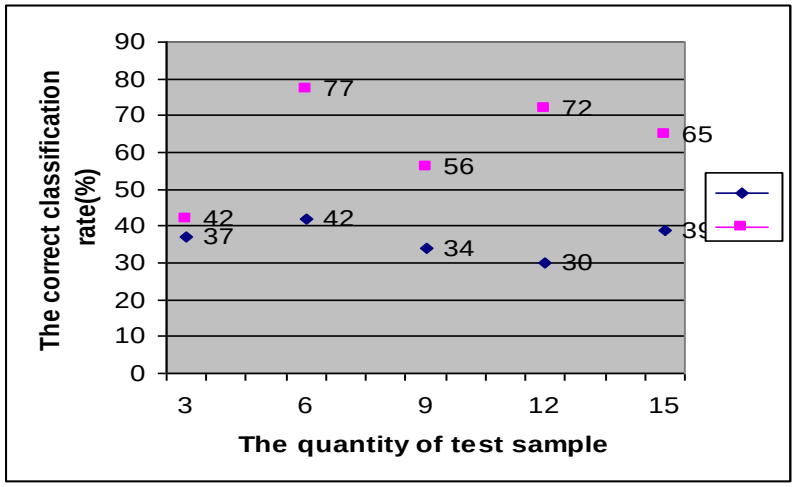


Fig. 8. Robustness test with 0.4 level salt and pepper noise

7. Analysis, Summary and Conclusion

The paper has described the procedure to be followed when designing an image classification system with either ANN's or SVM's. The binary classification structure of the SVM is extended by fusing multiple SVM's in order to classify multiclass data. The steps followed to pre-process the data for either the SVM or the ANN system is identical. To minimize computation burden, the dimensionality of image data is reduced with the HWT and PCA. This reduced dataset retains all the salient feature vectors which are applied to the classification system. The performance of the feed-forward ANN system was compared to that of the multiclass SVM with respect to generalization ability and robustness in the face of noise. The results were given in Fig. 6 to Fig. 8. From Fig. 6 we observe the following with regards to generalization: The SVM consistently outperforms the ANN irrespective of data size. For large quantities of data the performance of the ANN improves and reinforces the fact that ANN's require large quantities of data to produce an appreciable performance. With regards to robustness the following is observed: The SVM and ANN performance is comparable for small data quantities and same noise levels. The performance of the SVM deteriorates significantly for large data quantities. These results show that a well designed and trained ANN will have the inherent ability to consistently remain immune when faced with disturbances.

References

Chen C.H, Pau L.F., Wang P.S. (1993). "Handbook of Pattern Recognition and Computer Vision," World Scientific.

Cristianini N. and Shawe-Tayler J. (2000). "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods," Cambridge University Press.

David A., Lerner B. (2004-2005).Support vector machine-based image classification for genetic syndrome diagnosis. Pattern Recognition Letters, 26, 1029–1038.

Kilian, J., Siegelmann, H.K. (1996). The dynamic universality of sigmoidal neural networks. Information. and Computation, 128(1), 48-56.

Li. Z., Govender P., Kanny K., (2008). ANN based machine vision system for industrial sorting. "Proc.of the 2nd Robotics and Mechatronics Symposium," Bloemfontein, South Africa, Nov. 10-11,107-114.

Nixon, M.S.Aguado, A.S. (2008). "Feature Extraction and Image Processing", Academic Press.

Raviraj, P. Sanavullah, M.Y. (2007). The modified 2D Haar wavelet transformation in image Compression, Middle Easy Journal of Scientific Research, 2(2), 73-78.

Seetha, M., Muralikrishna, I.V., Deekshatulu, B.L., Malleswari, B.L., Nagaratna, Hegde, P. (2005-2008). Artificial neural networks and other methods of image classification. Journal of Theoretical and Applied Information Technology., 1039-1053.

Tomasi C., Manduchi R. (1998). Bilateral Filtering for Gray and Color Images, "Proceedings of the 1998 IEEE International Conference on Computer Vision".

Vapnik V.N. (1995). "The nature of statistical learning," Springer-Verlag.

Zhang, H., Cartwright, C.M, Ding, M.S. and Gillespie, W.A. (2000). Image feature extraction with various wavelet functions in a photorefractive joint transform correlator. Transactions on Optic Communications, 185, 277-284, Elsevier Science.