ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

EFFECTIVENESS OF PCA IN CLASSIFICATION PROBLEMS

Ankur Singh Bist
Computer Engineering, K.I.E.T, Ghaziabad, India

ABSTRACT

This paper is designed to make the study of effectiveness of PCA (Principal Component Analysis) method effectiveness in classification specially for the classification of face and computer virus. This technique is generally used for problems where eigenvector analysis is required. In order to analyze the varying trend in computer virology, the study of algorithms from different flavour and from various domains is required to make the vision clear towards viral detection.

KEYWORDS: Kits, Replication.

INTRODUCTION

There are various processes that have been used in the direction of classification of computer viruses from normal files that will finally lead to worm detection. Machine learning techniques are widely used in this direction. As statistics says that the attacks of malicious codes are increasing day by day so there is requirement of strong techniques that can be used for their detection. Malicious code designers use lot of techniques that are difficult to analyse and detect. The static methods also seems not to work in the case where every time there are rapid dynamicity from attacker side so now a days main focus is going on towards the methods that are dynamic and are able to detect zero day computer viruses [1]. The rise in the malicious threats like computer viruses activities are required to be handled and observed strongly to make certain defence that can stand as a saviour of security domain. Other types of malware are [3]:

1. Worms
2. Trojan horse
3. Botnets
4. Adware
5. Spyware

The mutating behaviour of metamorphic viruses is due to their adoption of code obfuscation techniques like dead code insertion, Variable Renaming, Break and join transformation. The given diagram shows the assembly file of the virus code.

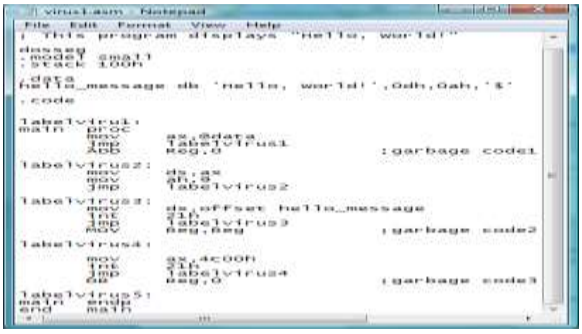


Figure 1: Assembly code of Virus File

VIRUS DETECTION METHODS

Computer viruses are growing very fast so in order to detect them various techniques are being used like detection using Hidden markov model, similarity analysis and other but the exactness and accuracy in the current scenario and for complex data set is an big issue.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

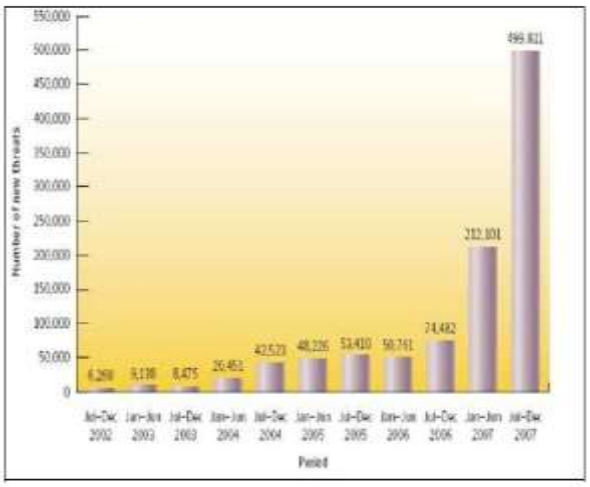


Figure 2: Malware Threat Rise

PCA is being used for face detection methods some other popular face classification methods are shown in diagram 2.

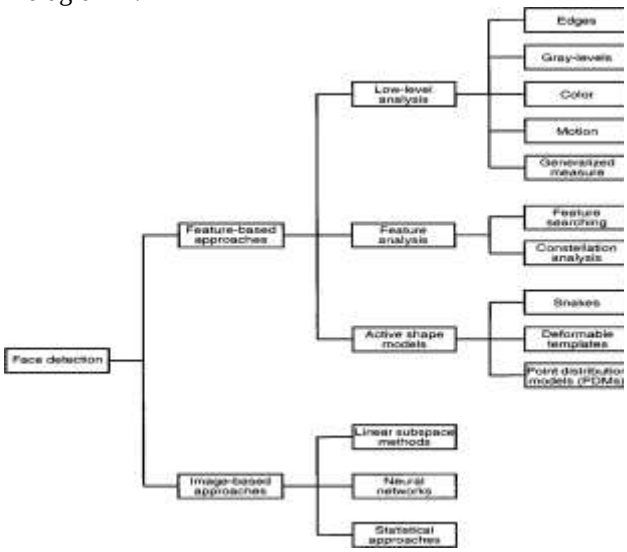
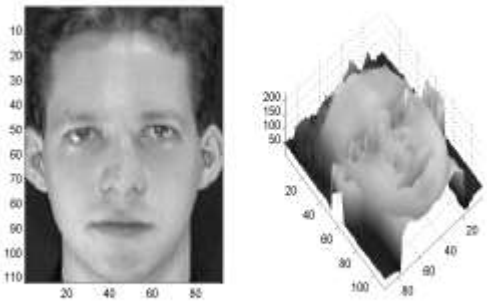


Figure 3: Techniques of Face Detection

The following diagram represents the dimensional view and eigen faces from dataset.



function pca (path, trainList, subDim) **INPUTS:**
path- full path to the normalised images from FERET database
TrainList- list of images to be used for training. names should be without extension and .pgm will be added automatically
subDim - Numer of dimensions to be retained (the desired subspace dimensionality). if this argument is ommited, maximum non-zero dimensions will be retained, i.e. (number of training images) - 1

OUTPUTS:
Function will generate and save to the disk the following outputs:
DATA- matrix where each column is one image reshaped into a vector - this matrix size is (number of pixels) x (number of images), uint8
imSpace- same as DATA but only images in the training set
psi- mean face (of training images)
zeroMeanSpace - mean face subtracted from each row in imSpace

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

pcaEigVals - eigenvalues
w - lower dimensional PCA subspace
pcaProj - all images projected onto a subDim-dimensional space

With the help of input parameters and output [4] that is to be obtained certain steps will be followed:-

1. Allocating memory for data matrix.
2. Creating data matrix
3. Creating training image space
4. Calculating mean face from training images
5. Dimension reduction
6. Subtract mean from all images
7. Project all images into lower dimension subspace

The methodology can be used for classification of computer virus. In the malware detection problem the "eigenfaces" termed as "eigenviruses". For making the analysis for viruses that leads to their classification the sequence of bytes are extracted from the computer virus files that are used for the training purpose. The scoring technique is to be used to classify the content. For making better analysis MWOR and NGVCK kit can be used to prepare the dataset.

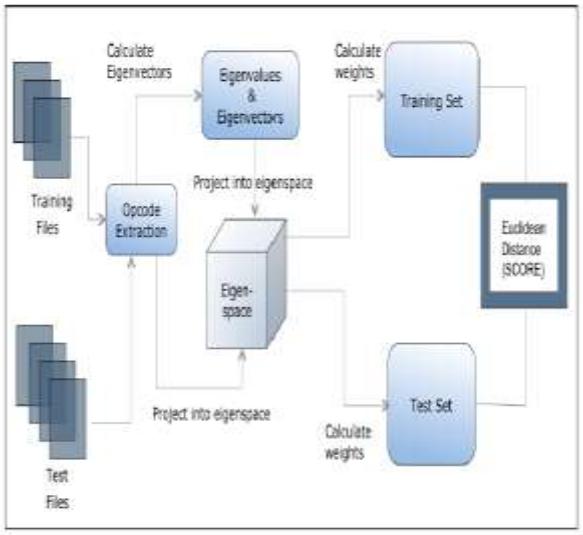


Figure 4: Process of classification [2]

Following evaluation specifications are used:-

a) **True positive-** It defines number of samples tested correctly and identified as virus code.

b) **True negative-** It defines the number of samples tested correctly and identified as benign executables.

c) **False positive-** It defines the number of sample tested and mistakenly classified as malicious executables.

d) **False negative -** It defines the number of sample tested and mistakenly classified as benign executables.

CONCLUSION

PCA is discussed in this paper in order to understand the effectiveness of its in various domain of classification in cases like computer virus and face detection it proved itself better. This study will be helpful for those working in the area of complex data classification.

ACKNOWLEDGEMENT

I wish to express my special thanks to all who supported me directly or indirectly in this work.

REFERENCES

1. www.wikipedia.com
2. **Deshpande, S. 2012.** Eigen value analysis for metamorphic virus detection Masters Thesis. Department of Computer Science, San Jose State University.
3. **VX Heavens.** http://vx.netlux.org/**Delac K., Grgic M., Grgic S.,** Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set,International Journal of Imaging Systems and Technology, Vol. 15, Issue 5, 2006.