A Statistical Approach to Malware Class Recognition

Aziz Makandar
Professor
Akkamahadevi Women's University, Vijayapura

Anita Patrot
Research Scholar
Akkamahadevi Women's University, Vijayapura

ABSTRACT
In this paper, we describe the proposed work on texture pattern classification using different Wavelet family, i.e. wavelet statistical features such as first order statistical feature vector. The WSF vector is formed to discriminate the various texture patterns of the Malware classes. The standard databases are used for experimental analysis of malware as a grayscale image. The database consists of 24 malware which belong to different variants with types of malware classes. The feature vector is further analyzed with malware classes the image to be classified based on the similarities in the image patterns. The experimental results shown that the efficiency of the wavelet based statistical features gives better classification results.

General Terms
Malware, Image Processing, Texture.

Keywords
Classification, Texture Pattern, Malware, Statistical Feature and Wavelet Transform.

1. INTRODUCTION
The analysis of texons played a major role in classification the pattern classification techniques and applications in the areas of image processing are growing increasingly. The image processing and pattern classification represents the state of art developments in the field. Texture pattern recognition is the task of classify input feature vector data in to classes based on the selected features from the vector. There are two types of classification supervised classification and unsupervised classification. The pattern recognition has applications in computer vision, SAR image classification, speech classification and texture classification. The texture classification plays a major role in many applications such as medical image analysis, pattern classification and so on. Supervised classification methods are used for face recognition, OCR, object detection and classification. Unsupervised classification methods are used in finding hidden structures, segmentation and clustering.

Wavelet transforms have become one of the most important and powerful tool of signal processing and representation. Now a day, it has been used in image processing, data compression and signal processing in different applications different wavelets are used. In this paper we present the overview of wavelets transformations in image processing. The objective of this paper is to give comparison results of the filter techniques with wavelet transformations.

Malware is software that performs unwanted features like Virus, Worm and Trojan horse. The functionalities of a malware such as execution and infection, self replication that infect another host, privilege escalation, manipulation that damages the host and concealment that hides from detection. The visualization of malware is an image is read as binary vector of 8 bit unsigned integers that are to be organized into a 2D array. This can be visualized as a gray scale image in the range [0, 255] the width of an image is fixed and height is allowed to vary depending on the file size.

2. RELATED WORK OF MALWARE
Texture plays a very important role in many research areas including image processing, pattern recognition, and medical image analysis also in computer vision. Texture analysis aims to finding a distinctive way of representing the primary characteristics of textures and represent them in some simpler but unique form, so that they can be used for robust, accurate classification and segmentation of objects. Through the texture statistical features plays a significant role in image analysis. Only a few architectures implement on-board textural feature extraction. Statistical texture features are formulated by using GLCM of malware image. The motivation of this work is that textures of a malware images are extracted effective features that considers the spatial relationship of pixels in a level co-occurrence matrix this matrix also called as gray level spatial dependence matrix a number of texture features may be extracted features namely contrast, correlation, energy and homogeneity are computed shows in table 2.

Texture analysis is a process of characterization of an image into different texture content. It is also an important research area in computer vision. Robert M Haralick.et.al [1][4] they introduce the method Gray Level Co-occurrence Matrix (GLCM) for images to extract matrix values are generated using this method is called Haralick features for classification. I. Buciu, and A. Gacsadi [2], these authors used the Gabor filters for feature extraction of medical images and classification. A. Eleyan, and H. Demirel [3], introduced the method for recognition of face by using GLCM features as statistical features for useful classification. T. Ojala.et.al [5], proposed a method for multi resolution approach for invariant texture classification for images. M. H. Bharati.et.al [6], proposed a comparison between various texture feature extraction methods as well as brief introduction. H.B.Kekre.et.al [7, 11], they proposed a method for content based image retrieval by using GLCM [1, 4]. Miroslav Benco.et.al [8], proposed a method for color image feature extraction by applying GLCM method on color image. Redouan Korchiyne.et.al [8-9] Dipankar Hazra.et.al [12] they proposed a method by combining wavelets, rotated wavelets and GLCM descriptors for extraction of features of images. Natraj.et.al [13] introduced the malware binaries as image which is having 8 bit vector of range 0 to 255 i.e. black and white as discussed in malware texture features. The malware images are looks like complete texture only that motivated to classify these images to particular class by applying classifiers[14-16].

3. PROPOSED WORK
3.1 Pre-processing
Pre-processing technique is an initial task of image processing. The texture pattern are discriminate the textures from the malware classes which belongs to the particular malware class. In this task we are normalizing the image in to

256X256 ranges which is in gray scale image. Further this input pattern is send to the pre-processing.

3.2 Feature Extraction

A co-occurrence matrix is also called as distribution of matrix, distribution that is defined as distribution of co-occurring values at a given offset. The mathematically a co-occurrence matrix C is defined over an n x m image 'I', parameterized by an offset $\Delta x, \Delta y$ where 'i' and 'j' are the image intensity value, p and q are the spatial positions in the image i and the offset depends on the direction used θ and the distance at which the matrix is computed 'd'.

The co-occurrence matrix can measure the texture of the image because co-occurrence matrices are typically large and sparse; various metrics of the matrix taken to get a more useful set of feature these features generated using methods are called Haralick features [1]. The co-occurrence matrices captures numerical features of a texture using spatial domain relation of similar gray tones numerical features computed from the co-occurrence matrix can be used to represent, compare and classify textures [2]. The value of an image is referred as grayscale value of specified pixel or intensity.

$$C_{\Delta x,\Delta y}(i,j)\sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, \text{if } I(p,q)=i \text{ and } I(p+\Delta x, q+\Delta y)=j\\ 0, \quad \text{otherwise}\end{cases} \quad (1)$$

The following is a subset of standard features derivable from a normalized co-occurrence matrix. The matrix is the number of rows and columns is equal to the number of levels G in the image the matrix element.

P(i,j|Δx,Δy) is the relative frequency with two pixels separated by a pixel distance Δx,Δy occur within a given neighborhood, one with intensity 'i ' and 'j'. The matrix element P (i,j|d,ɵ) contains the second order statistical probability values for changes between gray levels 'i' & 'j' at a particular displacement distance d and at a particular angle(ɵ). Using a large number of intensity levels G implies storing a lot of temporary data, i.e. a G × G matrix for each combination of (Δx, Δy) or (d,ɵ). Due to their large dimensionality, the GLCM's are very responsive to the amount of the texture samples on which they are predictable.

Energy: $\sum_i \sum_j p[i,j]^2$ (2)

Contrast: $\sum_{n=0}^{Ng-1} n^2 \left\{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p[i,j]\right\}$ (3)

Correlation: $\frac{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(i,j)p[i,j]-\mu_x\mu_y}{\sigma_x\sigma_y}$ (4)

Entropy: $\sum_i \sum_j p[i,j]\log(p[i,j])$ (5)

The table illustrates the texture features of GLCM approach for an texture image as well as mathematical representation of each feature are equations such as (1)(2)(3)(4) and (5). The level of co-occurrence matrix (GLCM) [2] methods are a way of extracting second order statistical texture features. This advance has been used in a number of applications. These are theoretically possible but not commonly implemented due to calculation time and interpretation difficulty.

3.3 Classification

It identifies the subset of Malwares by preserving only the most important predictors and filtering or excluding all others. The scale and translation parameters are given by,S=2-m and T=n2-m where m ,n are the subset of all integers. Thus, the family of wavelet is defined in equation 6.

$$\psi_{m,n}(t)=2^{\frac{m}{2}}\psi(2^m t-n) \quad (6)$$

The wavelet transform decomposes a signal x(t) into a family of waveletsas given in equation 7 and

$$x(t)=\sum_m\sum_n C_{m,n}\,\psi_{m,n(t)} \quad (7)$$

Where, $C_{m,n}=\langle x(t),\psi_{m,n}|(t)\rangle$ (8)

For a discrete time signal x[n], the decomposition is given by:

$$x_{[n]}=\sum_{i=1td}\sum_{k\epsilon z} C_{i,k,g[n-2^i k]}+\sum_{k\epsilon z} d_{l,k\,[n-2^i k]} \quad (9)$$

In case of images, the DWT is applied to each dimensionality separately. The resulting image X is decomposed in first level is xA, xH,xV and xD as approximation, horizontal, Vertical and diagonal respectively. The xA component contains low frequency components and remaining contains high frequency components. Hence, X= xA+{xH+xV+xD}. Then DWT applied to xA for second level, third level and fourth level decomposition. Hence the wavelet provides hierarchical framework to interpret the image information. Wavelet transform that is localized on mother wavelet, the Statistical Feature Extraction (SFE) stage we are applying wavelet filters such as Discrete Wavelet Transform then the extracted 11 statistical features are constructed a feature vector and to get normalized features for classification.

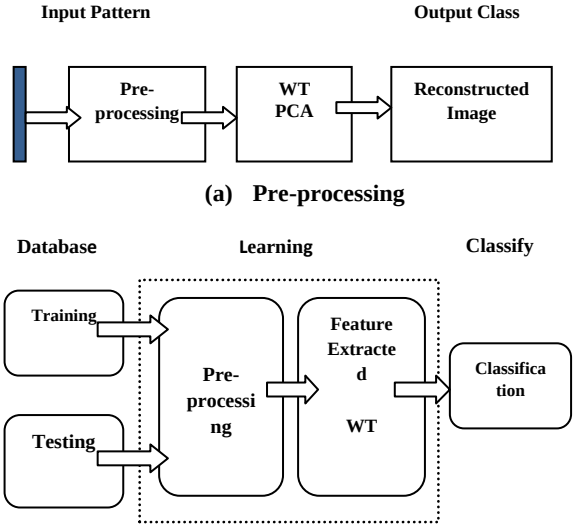

Fig 1: Proposed Algorithm

4. EXPERIMENTAL ANALYSIS

The result analysis is done on the malware dataset which consists of 24 malware family with 3131 malware samples. Each family of malware consists of 80 to 300 or more samples. The texture patterns of the malware samples are similar that motivates to classify the malware samples based on the statistical approach and wavelet transform. The wavelet transform done on the input samples and decomposition is done after that principal component analysis is used to reduce the dimensionality of the decomposed samples. To calculate the statistical features such as mean, mode, standard deviation, correlation, contrast, entropy, energy, RMS and variance. The build feature vector is used to discriminate the texture patterns from the malware samples and classify them as shown in fallowing fig.

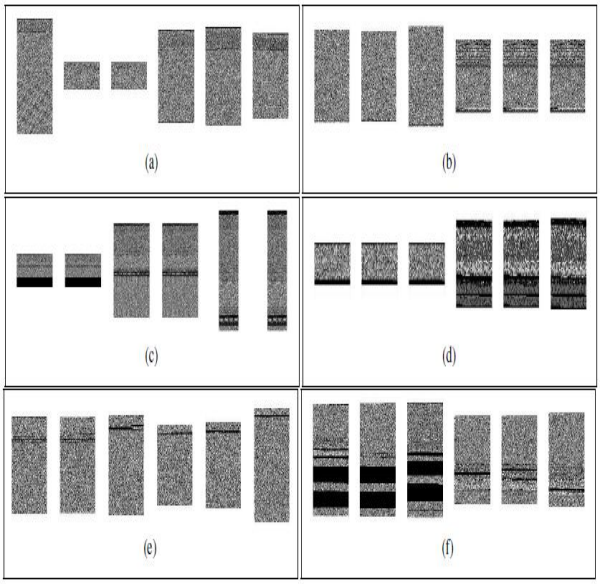

Fig 2: Malware Classes (a) Allaple (b)Ejik (c) Mydoom (d) Tibs (e) Udr. and (f) Virut.

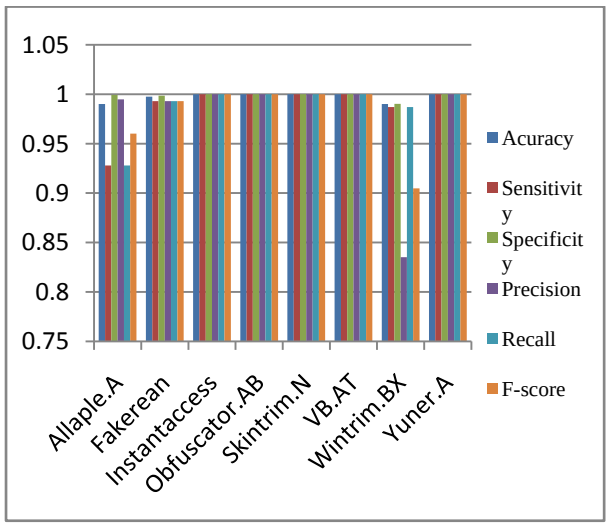

Fig 3: Malware Classification measures Accuracy, Sensitivity, Specificity, Precision, Recall, F-score.

Table 1. Malware Class with measures

Malware Class	Accuracy	Sensitivity	Specificity
Allaple.A	99.0%	92.7%	99.92%
Fakerean	99.75%	99.29%	99.84%
Instantaccess	100%	100%	100%
Obfuscator. AB	100%	100%	100%
Skintrim.N	100%	100%	100%
VB.AT	100%	100%	100%
Wintrim.BX	99.00%	98.70%	99.02%
Yuner.A	100%	100%	100%

Table 2. Malware Class measures

Malware Class	Precision	Recall	F-score
Allaple.A	99.48%	927%	96.01%
Fakerean	99.29%	99.2%	99.29%
Instantaccess	100%	100%	100%
Obfuscator. AB	100%	100%	100%
Skintrim.N	100%	100%	100%
VB.AT	100%	100%	100%
Wintrim.BX	83.51%	98.70%	90.47%
Yuner.A	100%	100%	100%

Table 3. Classification Measures

Classification	Accuracy	Sensitivity	Specificity
KNN k=3	99.72%	98.84%	99.84%

Table 3. Classification Measures

Classification	Precision	Recall	F-score
KNN k=3	97.7%	98.84%	98.22%

5. CONCLUSION & FUTUREWORK

We proposed an efficient malware class recognition technique based on texture of malware variants. In this paper proposed an approach for malware texture patterns are extracted by GLCM approach and build a feature vector. That is used for extracting second order statistical texture parameters. The KNN classifier with k=3 gives 99.72% of accuracy. Enormous efforts have been made in search of an efficient texture description and texture analysis. It is generally a difficult problem due to diversity and complexity of natural textures and these features are useful in classification of images as well as used for further classification. The future work is used for unsupervised learning classification of malware variants. These are real time pattern recognition applications like Military & Medical Applications.

6. ACKNOWLEDGMENTS

This research work is funded and supported by UGC under Rajiv Gandhi National Fellowship (RGNF) UGC Letter No: F1-17.1/2014-15/RGNF-2014-15-SC-KAR-69608, February, 2015.

7. REFERENCES

[1] Robert. M Haralick, K Shanmugam, Dinstein. “Textural Features for Image Classification,” IEEE Transactions on Systems, Man, and Cybernetics. pp. 610–621, 1973.

[2] I. Buciu, and A. Gacsadi, “Gabor wavelet based features for medical image analysis and classification,” IEEE 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies, pp. 24-27, 2009.

[3] A. Eleyan, H. Demirel, "Co-Occurrence based Statistical Approach for Face Recognition", Computer and Information Sciences, 2009.

[4] R. M. HARALICK, "Statistical and structural approaches to texture", Proc. IEEE, pp. 786-804, 1979.

[5] Aziz Makandar and Anita Patrot,"Malware Image Analysis and Classification using Support Vector Machine ," International Journal of Trends in Computer Science and Engineering, Vol.4, No.5,pp.01-03,2015.

[6] T. Ojala, M. Pietikainen, T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns", IEEE. Trans. On Pattern analysis and Machine intelligence, 2004.

[7] M. H. Bharati, J. Liu, J. F. Mac Gregor, "Image Texture Analysis: methods and comparisons", Chemometrics and Intelligent Laboratory Systems, pp. 57- 71, 2004.

[8] H.B.Kekre, Sudeep D. Thepade, Tanuja K. Sarode and Vashali Suryawanshi , "Image Texture Feature Extraction Using GLCM Approach",International Journal of Scientific and Research Publications, Volume 3, Issue 5, ISSN 2250-3153, May 2013.

[9] Martina Zachariasova, Slavomir Matuska, Kamencay, "An Advanced Approach to Extraction of Colour Texture Features Based on GLCM", International Journal Advance Robotic System, doi: 10.5772/58692, 2014.

[10] Redouan Korchiyne, Sidi Mohamed Farssi, Abderrahmane Sbihi, Rajaa Touahni, Mustapha Tahiri Alaoui., "A combined method of Fractal and GLCM features for MRI and CT scan Images Classification," Signal & Image Processing an International Journal (SIPIJ) Vol.5, No.4, August 2014.

[11] Dipankar Hazra," Texture Recognition with combined GLCM, wavelet and Rotated wavelet Features.", International Journal of Computer and Electrical Engineering, Vol.3, No.1, pp.1793-8163, February, 2011.

[12] Aziz Makandar and Anita Patrot, "Computation Pre-Processing Techniques for Image Restoration," International Journal of Computer Applications (0975-8887), Volume 113, No.4,pp.11-17, March 2015.

[13] H.B. Kekre, Sudeep D. Thepade, Tanuja K. Sarode and Vashali Suryawanshi," Image Retrieval using Texture Features extracted from GLCM, LBG and KPE", International Journal of Computer Theory and Engineering, Vol.2,No. 5, ISSN 1793-8201, October, 2010.

[14] Nataraj. L. Karthikeyan. S, Jacob, G. and Manjunath. B. "Malware Images: Visualization and Automatic Classification," Proceedings of the 8th International Symposium on Visualization for Cyber Security, Article No. 4. 2011.

[15] Aziz Makandar and Anita Patrot," Review on malware analysis and detection," International Journal of Computer Applications (0975-8887) National Conference on Knowledge Innovation in Technology and Engineering NCKITE 2015, pp.35-40.

[16] Aziz Makandar and Anita Patrot, "Color Image Analysis and Contrast Stretching using Histogram Equalization," International Journal of Advanced Information Science and Technology (IJAIST) ISSN 2319:2682, Vol.27, No.27, July 2014,pp.119-125.