An Exhaustive Survey on Malware Classification

Abhijeet Singh Sran and Ankur Singh Bist
Dept. of Computer science and engineering
KIET group of institutions, Ghaziabad

Abstract

We present a very simple and an effective survey for understanding and classifying malware using various image processing techniques. Malware files are treated as grayscale images, and the images belonging to the same family appear very similar in many aspects such as texture, layout and design. By considering this visual similarity a classification method is put forward. Previous experiments show results that are very accurate with more than 90% classification accuracy on a malware database containing various samples with more than 25 different families.

Keywords

Malware visualisation, Image Texture, Image processing.

Introduction

Traditional way of analysing malware involves extraction of binary signatures including their fingerprint. Due to the rapid spread of malware, there is an exponential increase in the number of new signatures that are released every year. The other approach of malware analysis includes static code analysis and dynamic code analysis. In static code analysis the code is broken and the control flow of the executable file is explored to look for malicious patterns. On the other hand, in Dynamic analysis a virtual environment is needed for executing the code and a report is generated based on its execution trace. There are advantages as well as disadvantages of both these techniques. Complete code coverage is offered in static analysis but it suffers from the problem that the code has to be first unpacked and decrypted before analysis. The dynamic approach is more efficient and the executable code do not need to be unpacked or decrypted before analysis. However, it is more time and resource consuming and has scalability issues.

Malware Detection Technique

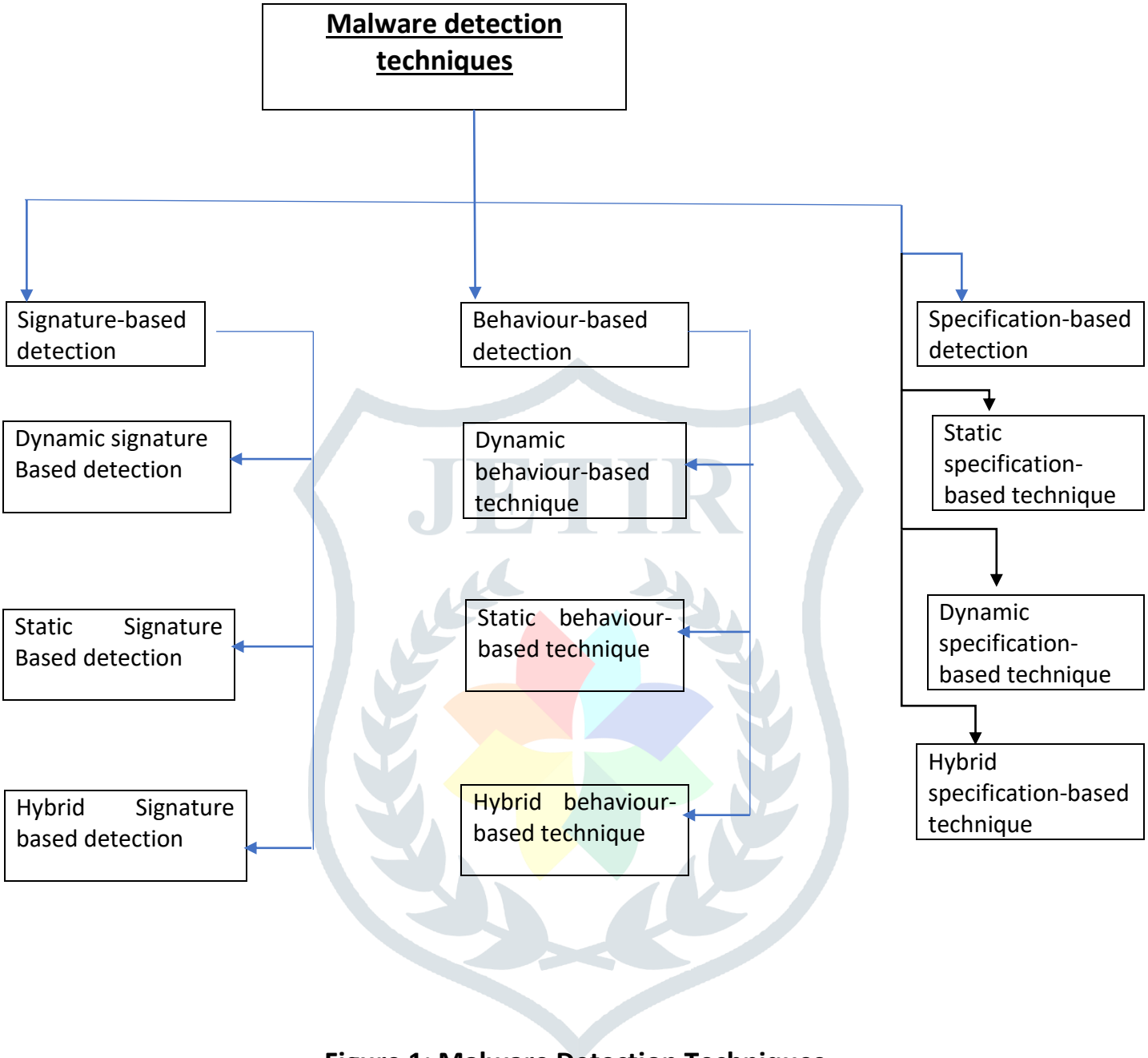

Figure 1: Malware Detection Techniques

Related Work

Binary data can be visualised and manipulated by using several text editors and binary editors. There have also been several GUI based tools which help to compare the files. Yoo[7] used self-organized maps that are used to detect and visualise malicious code inside anexecutable file. Quist and Liebrock [11] developed a visualization framework for reverse engineering. They identified functional areas through a node link visualization where the

addresses are represented by nodes and the addressed are represented by the link between nodes.

While there hasn't been much work on viewing malware as digital images, Conti et al. [8,9] showed that they can automatically classify the different binary fragments using statistical features. However, they are only concerned about primitive binary fragments and not malware. They work by presenting the malware as a grayscale image. There are several techniques for classification of malware. These include static analysis as well as dynamic analysis. In Rieck et al. [10] the behaviour of the malware is checked before classifying it into the malware family. A labelled dataset of more then 10000 malware samples were labelled by an antivirus software and were divided in to 14 families. Then they checked the behaviour of all the malware in a sandbox environment and a behaviour report was generated. From that report they can generate a feature vector for every malware.

Visualization

If we are given a malware then it can be read as a vector of 8-bit unsigned integers and then organized into a 2D array. This can then be visualized as a grayscale image in the range [0,255].

The image has a fixed width and the height varies according to the file size as shown in figure 2.

Malware binary ------->binary to ----------> 8-bit vector to -------> grayscale

01110011100100001….8-bit vector grayscale image image

Figure 2: Malware binary to grayscale image

Malware classification

An observation showed that images of different malware samples from a given family appear visually similar and distinct from those belonging to different family [1]. This similarity led us to look at malware classification using techniques from computer vision.

Malware can be broken down into the following classification [6].

Viruses: A computer virus takes the control of functions on your computer and might delete some data and try to capture the personal information.

Rootkit: A rootkit is a virus in complete stealth mode and is very difficult to find and remove.

Spyware: It does what its name implies. It spies on you. It watches what website you visit, reports to the central server and then serves up advertisements based on your surfing content.

Worms: Worms are meant to propagate and infect as many computers in any short span of time. These are often spread by email sending themselves out to every person in your contact list without your knowledge.

Trojan Horse: A Phishing scam is a technique used to get a person either through email or a website to input data into fields they believe are legitimate.

Adware: An adware is a software that automatically displays and downloads advertising materials such as banner ads and pop up ads.

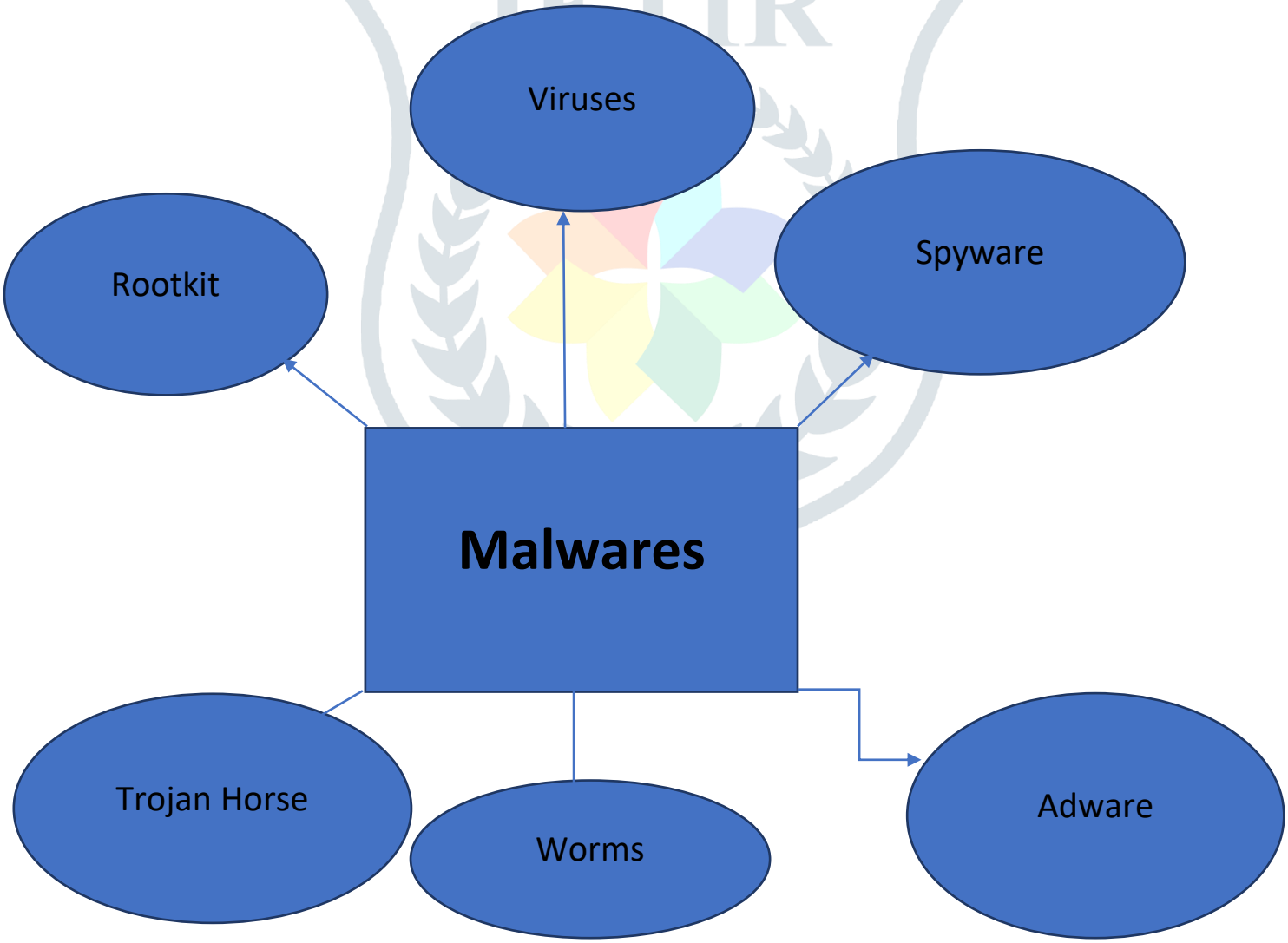

Figure 3: Basic Malware Classification

Feature vector and classifier

There are several features for the analysis of malware and virus. One of the most common method is to analyse the frequency of a texture block. In standard approach the frequency domain is divided into rings and wedges and the features are computed in these regions [3,4]. Gabor filtering is a popular computational approach. A Gabor filter is a filter that is frequency and orientation selective. By varying the frequencies and orientations, we obtain a bank of Gabor filter. Several filtered images are obtained when the image is passed through this filter and from that filtered images the texture- based features are extracted. Texture features using Gabor filter have been successful in texture classification and segmentation. This process is very helpful for malware detection. [2,5].

Limitations

Although an image processing approach is a good and a popular approach to analyse malware, still there are some countermeasures that can beat the current technique as it is based on global image-based features. More local feature extraction schemes that consider the distinct characteristics of malware executables and their primitive binary segments are being explored to tackle such attacks. Segmenting out the image regions and characterizing the local texture and spatial distribution of these texture patters is a good future extension to prevent against attacks.

Challenges in malware detection

The different challenges in malware detection are listed below[6]

Performance

- Detection is not as fast enough to be used in a browser.

Accuracy

- False positive rates of 5% is acceptable for static analysis tool but is over 100x what is acceptable for in-browser detection.

Obfuscated malware

- Most of the JavaScript codes are frequently unclear for static detection and are hence ineffective.

Malware transience

- Only offline is not effective because web malware “infects and dies young”.

Conclusion

In this paper asurvey is presented for malware analysis based on image processing and visualizing techniques. Malware is characterized globally by characterizing a common image feature descriptor. Theobservations are very encouraging and the classification accuracy is also very high with image based features and is competitive with the state of art results in the literature at a significantly less cost. Analysis of malware using computer vision techniques gives us a broader spectrum of viewing and analysing malware.

References

1) Bist, Ankur Singh. "Detection of metamorphic viruses: A survey." Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE, 2014.

2) Wei-Chung Huang, Fabio Di Troia and Mark Stamp Department of Computer Science, San Jose State University, San Jose, California, U.S.A.

3) L. Nataraj, S. Karthikeyan, Dept. of Electrical and Computer Engineering, University of California.

4) G. Jacob, Dept. of Computer Science, University of California, Santa Barbara.

5) B. S. Manjunath Dept. of Electrical and Computer Engineering, University of California, Santa Barbara.

6) Jyoti Landage, Me, Dept. of Comp Engg Sinhgad College of Engg, Vadgaon, Pune

7) Yoo, I. Visualizing Windows Executable Viruses Using SelfOrganizing Maps., 2004 International Workshop on Visualization for Cyber Security (VizSec).

8) Conti, G. and Bratus, S. 2010. Voyage of the Reverser: A Visual Study of Binary Species, Black Hat USA.

9) Conti, G. Bratus, S. Shubina, A. Lichtenberg, A. Ragsdale, R. Perez-Alemany, R. Sangster, B. and Supan, M. 2010. A Visual Study of Binary Fragment Types Black Hat USA.

10) Rieck, K. Holz, T. Willems, C. Dussel, P. and Laskov, P. Learning and classification of malware behaviour. 2008. Fifth Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA'08), pages 108–125.

11) Quist, D.A. and Liebrock, L.M. 2009. Visualizing compiled executables for malware analysis. International Workshop on Visualization for Cyber Security (VizSec), 27-32