IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 2, Issue 11, November 2014 ISSN 2321-5992

A Novel Approach to Digital Forensics Image Mining System

[1]Onyemauche U.C.,[2]Okonkwo O.R.

Department of Computer Science
Nnamdi Azikiwe University Awka Anambra State, Nigeria

ABSTRACT

There has been an escalating amount of illegal image data transmitted via the internet. This has prompted the call to develop effective image mining systems for digital forensics purposes. This paper discusses the requirements of digital image forensics which emphasize the design of our forensic image mining system. This system can be trained by a hierarchical Support Vector Machine (SVM) to detect objects and scenes which are made up of components under spatial or non-spatial constraints. Forensic investigators can communicate with the system via a grammar which allows object description for training, searching, querying and relevance feedback. In addition, we propose to use an AI based networks approach to deal with information uncertainties which are inherent in forensic work. An analysis of the performance of the first prototype of the system is also provided.

Key words:- Artificial Intelligence, Image mining, timestamp, Event correlation.

1. INTRODUCTION

Image mining is part of the interdisciplinary field of knowledge discovery in databases[1]. Digital forensics is the branch of forensic involving the recovery and investigation of material found in digital devices due to incident of computer crime occurrence. Image mining is part of the interdisciplinary field of knowledge discovery in databases. Digital forensic is a synonym for the computer forensic in early start but today it includes other area of investigation like computer, database, and network, mobile which are capable of storing digital data[2]. Techniques used for such investigations are varied and may include data mining and analysis, event correlation, information hiding analysis, etc. Since multimedia format is widely used and readily available via the Internet, there are increasing criminal activities in the last few years, which involve the transmission and usage of inappropriate material such as child pornography in this format. Hence, much forensic evidence comes in the form of images or videos that contain objects and/or scenes that may be related to criminal behaviours. A typical investigation in digital forensics can generate large image and video data sets. For example, a disk can easily store several thousands of images and videos in normal files, browser cache files and unallocated space (i.e., non-file system areas on the disk which may contain fragments of files). This can make the task of searching for, and retrieving, images/videos very time consuming. Digital Image Forensics (DIF) efficiently seeks for evidence by using appropriate techniques based on image analysis, retrieval and mining. Owing to rising criminal activities through the use of digital materials, the use of such techniques for investigative purposes have only recently emerged, although they have been intensively researched over the last three decades for many other important applications such as medical diagnosis, mineral exploration, environmental monitoring and planning, aerial surveillance to mention but a few. When the digital forensic is in consideration usually three different sets of people from Law Enforcement agencies, Military, Business & Industry are involved with the intention of tracking down criminals who attack the security of systems and use computers for unauthorized activities. Digital Forensic address the issues of National and Information Security, Corporate Espionage, White Collar Crime, Child Pornography, Traditional Crime, Incident Response, Employee Monitoring, Privacy Issues. Since the quality of the retrieval results relies on the choice of features and their similarity measures, much research has been focused on identifying features with strong discriminatory power and similarity measures that are meaningful and useful. In addition, we would ideally want a more “intelligent” system which can include high-level knowledge, deal with incomplete and/or uncertain information, and learn from previous experience. Such systems could include, for example.

- Model-based Methods: A model of each object to be recognized is developed. These objects are classified using their constituent components that in turn are characterized in terms of their primitives,
- Statistical Modeling Methods: Statistical techniques are used to assign semantic classes to different regions/objects of an image, and − User Relevance Feedback Methods: User feedback is required to drive and refine the retrieval

process. The system is thus able to derive improved rules from the feedback and consequently generate better semantic classes of images.

Model-based methods exploit detailed knowledge about the object and are capable of reasoning about the nature of the object. Moreover, the models created are often handcrafted and cannot easily improve their performance by learning. Statistical modeling techniques rely on statistical associations between image semantics and, as such, do not require the generation of any complex object model. Such associations can be learned using the statistical model. When the digital forensic is in consideration usually three different sets of people from Law Enforcement agencies, Military, Business & Industry are involved with the intention of tracking down criminals who attack the security of systems and use computers for unauthorized activities. Digital Forensic address the issues of National and Information Security, Corporate Espionage, White Collar Crime, Child Pornography, Traditional Crime, Incident Response, Employee Monitoring, Privacy Issues[3].

2.THE CHALLENGES FACING COMPUTER FORENSICS INVESTIGATORS IN OBTAINING INFORMATION FROM MOBILE DEVICES FOR USE IN CRIMINAL INVESTIGATIONS

There are a number of electronic personal devices that are labeled mobile devices" on the market today. Mobile devices include cell phones; smart phones like the Apple iPhone and Blackberry; personal digital assistants (PDAs); and digital audio players such as iPods and other MP3 type devices. Laptop computers, tablets and iPad products are not typically classified as a mobile device as they are not small enough to be considered handheld. Today, the ever popular smart phone comes with a storage capacity that is similar to a laptop while commonly utilized as a portable office, social network and entertainment center all rolled into a solitary, convenient device. A smartphone is a mobile device that provides advanced computing and offers the ability to run mobile applications with more connectivity options than a cellular phone[4]. Technological and storage capacity for mobile devices has grown exponentially. Over the last decade, capabilities and features of mobile devices have turned them into data repositories that can store a large amount of both personal and organizational information. Unfortunately, criminals have not missed the mobile device information revolution. Within the past few years, they have increasingly been using mobile phones and other handheld devices in the course of committing criminal acts. For example, a drug dealer may keep a list of customers who owe him money in a file stored on his handheld device or a child pornographer could keep nude images of underage children engaging in sexual activities on a mobile device for the purposes of trading photos or video files with other pedophiles. Indeed, almost every class of crime can involve some type of digital evidence from a device that is essentially a portable data carrier. This increases the potential for incriminating data to be stored on mobile devices and to be utilized as evidence in criminal cases. Can valuable information be obtained from a mobile device to assist in a criminal investigation? What are the challenges a forensics investigator faces in obtaining information from these devices? Mobile devices can contain such electronic records information such as electronic mail, word processing files, spreadsheets, text messages, global positioning system (GPS) tracking information and photographic images that can provide law enforcement personnel with essential evidence in a criminal investigation[5]. A mobile phone's ability to store, view and print electronic documents is easily utilized from a single hand-held device with the processing power and the storage capacity similar to a bulky laptop.

3. OPERATIONAL MODEL

In order to design and implement an efficient image mining system architecture, an operational model of the digital forensic image mining process was developed. This model reflects the procedures undertaken by an investigator during a typical digital forensics investigation. The model consists of two "activities", namely one involving the rapid reduction of the large quantity of evidence that is involved in a case, and one involving the core image mining activities that deal with the actual image retrieval process for digital forensic examination. The former activity, as mentioned later, involves the execution of a chain (in reality, a forest of connected trees) of forensic tools for analyzing the content of large data streams (disks and other data), filtering the data streams for data reduction, extracting meta-data (eg, file timestamps) etc. to downstream analysis and decision making that leads to a successful investigation. The latter activity is simply one of the many possible forensic tools deployed in the case investigation graph. The core image mining operational model follows two stages, namely the training phase and the testing or classification phase. The training phase, also referred to as the classification model-generation phase, builds the object models relevant to the particular domain at hand. This phase is usually undertaken by an experienced investigator who has an insight into the object types involved in the particular case under investigation, an understanding of the classifier, knowledge of the object layout (eg, constraints such as positions, orientations etc.) and so on. The investigator will also be responsible for providing the relevance feedback on a priori evidence (eg, images from similar cases) in order to refine and improve

the quality of the classification model. We propose to use a Event Correlation for query refinement with a set of relevance feedback parameters The testing phase uses the refined classification model (given by the set of model parameters) developed during the training phase to classify the set of images found in the case under investigation.
We have designed and developed a complete operational system for digital forensics which implements both the digital forensic examination process (the chain of forensic tools) as well as a prototype model-building and classifier system that focuses on the core image mining component of the operational model. In this paper, we focus on the model-building and classifier system.

3.1 Recognition of Component-based Objects and Scenes

There have been various component-based systems which deal with human detection. For example, features such as eye, nose, and mouth are first detected and then combined in a spatially constraint configuration in order to determine a face e.g.[6] Other systems detect humans and their actions for various purposes: surveillance (e.g. detection of criminal activities; movement recognition (e.g. gesture recognition for interactive dance systems. The underlying models for such methods can be grouped in two main categories: task-specific models and general models that can be applied to specific tasks. The task-specific approach constructs a model from the components of a human silhouette and tightly coupled it with constraints that govern a specific action of interest. This approach is rather restrictive and does not provide a framework that can be readily extended in order to model different behaviors for other applications. The general approach, on the other hand, constructs a model from primitives in a bottom-up fashion and uses a regular grammar to represent various modes of motion and interactions e.g. The system is then trained using models that represent certain exemplar behaviours. A special type of statistical models called Hidden Markov Models (HMM) is used to represent both a priori knowledge and new knowledge resulted from new behaviours. Low level primitives are firstly detected before they are passed into the grammar for behaviour analysis. These systems, although robust, rely on motion information to resolve ambiguities. We extend the approach by which used Haar wavelet coefficients as features and SVMs (Support Vector Machine) for training. In their system, the magnitude of the coefficients of two scales (16x16 and 8x8 pixels) and three orientations (horizontal, vertical, diagonal) that indicate the intensity variation are used to locate the position of the components of objects. This multi-level approach is robust and flexible for object configuration design. One drawback is that difficulties due to image scaling and transformations have not been addressed. Our image mining system for computer forensic purposes allows the use of other features (e.g. texture features) in addition to Haar coefficients. We also investigate the effects of using different colour spaces, and of image scaling and transformations. In addition, we examine the needs of effective communication and usage of the system by forensic investigators and relevance feedback for continuous improvement. To this end, we develop a grammar to facilitate the specification of objects, scenes and their relationships. This grammar can also help to filter out invalid configurations. Relevance feedback will be provided.

4. RESULTS

One application that can benefit from our image mining system is to detect and filter out improper images such as those of partially clad people. We use this application as a case study to test the performance of this system. We use a training set of 200 images consisting of 104 positive images of partially-clad people, and 96 images of negative images of landscapes, textures, clothed people, sport scenes, etc. The patch detectors firstly detect face, waist and pelvis; then combine these components into a hierarchy to detect partially-clad people. Each feature vector is composed of high edge coefficients defining the outline of body parts and regions of continuous tones (e.g. bare skin, texture, colour). We perform three experiments using different colour spaces and varying the use of texture homogeneity values. The first test uses HSV space, maximum value of wavelet coefficients in Hue and Value as edge coefficients, and the variance of Hue and Saturation for homogeneous regions. 92% true positive and 74% true negative detection rates are obtained.
The second test uses YCbCr space, maximum values of Cb and Cr, and the variance of Cb and Cr. 79% true positive and 95% true negative detection rates are obtained. The third test is similar to the second test except that texture homogeneity values are included as features instead of the variances. The detection rates are the same as in the second test.

4.1 Discussion

From these results, we have found that HSV is more useful for finding positive images, while YCbCr is more discriminating but at a reduced rate of positive detection. The texture homogeneity is not a discriminating feature for this application. Interestingly, we observed that the skin detection using YCbCr has a similar positive rate to that of the SVM classifier. Does this imply that the rate of improvement rests with the choice of a better colour model for skin detection? To facilitate the communication between forensic investigators and the system, we develop a grammar for describing objects and scenes as hierarchies of component detectors. This grammar defines the position, orientation, error bound, and varying levels of resolution, to allow fast search of regions of interest and more detailed and computationally expensive search at a finer level. Users can use this grammar for three tasks: to specify objects and

scenes for training, for querying and for providing feedback to the system. Information on the position and orientation is expressed in numerical quantities, while relative spatial arrangement can be expressed in either absolute measurements, or precise terms (e.g. north, south, east, west), or fuzzy terms (e.g. up, down, above, below). These hierarchies which can be represented in tree data structure are encapsulated into a file grammar to support storage and manipulation for future use.[7]

4.2 Forensic-Scene:Algorithm

Scene
Scene-Detector-ID
Comp-Detector-ID
End-Scene
Scene
Scene-Detector-ID
Object-Detector-ID
End-Scene
Comp-Detector:
Component
Comp-Detector-ID
Comp-Detector-Loc
Displacementopt
Orientationopt
Relation-Listopt
End-Component
Object-Detector:
Object
Object-Detector-ID
Object-Detector-Loc
Displacementopt
Orientationopt
Relation-Listopt
Detector-List
End-Object
Detector-List:
Detector-List, Gen-Detector-ID
Gen-Detector-ID: one of
Object-Detector-ID,
Comp-Detector-ID,

Scene-Detector-ID

Object-Detector:
Object
Object-Detector-ID
Object-Detector-Loc
Displacementopt
Orientationopt
Relation-Listopt
Detector-List
End-Object
Detector-List:
Detector-List, Gen-Detector-ID
Gen-Detector-ID: one of
Object-Detector-ID,
Comp-Detector-ID,

Scene-Detector-ID

4.2.1 Explanation

This grammar is extensible to include non-spatial relationships and dynamic scenes. Non-spatial relationships would allow users to specify special characteristics of image evidence based on their previous experience. For example, the co-occurrence of bare skin and pixilated image regions might heighten the chance that the image is pornographic; the co-occurrence of weapons and important buildings might indicate a breach of security. Dynamic scenes occur in motion videos when objects may appear or disappear, or the attributes and relationships between objects may change. These changes can be implemented by appropriate operations on the tree (insertion, deletion, modification of attributes in the node contents by traversing the tree). Standard transformations (scale, translate, rotate, shear) and linguistic modifications of spatial relationships may be treated as changes in object attributes. To track an object that may be occluded from time to time, a visibility flag is used.

5. CONCLUSION

The presented image mining system provides the facility for training the system to detect the image evidence required, as well as for correcting inaccurate search results or fine-tuning the search further. The communication between users and the system is facilitated by an adaptive grammar[8]. To date, the prototype system consisting of the component-based detection engine and the grammar has been implemented and evaluated for detection of images containing partially clad humans and in other applications with very promising results. The system architecture is flexible in the sense that other types of classifiers (e.g. Naïve Bayes, C4.5 or neural networks) can be used instead of the SVM if they are more suited to the classification of specific types of data. Moreover, different classifiers may be used for different parts of the system. The grammar is generic and extensible to allow more sophisticated query to be generated if required[9]. Event Correlation rules provide a compact and efficient means to represent joint distributions over a large number of random variables and allows effective inference from observations (e.g. [10]). Hence, they can be used to understand and learn probabilistic and causal relationships through updating beliefs based on evidence provided. The need for dealing with uncertainties that are inherent in Digital Forensics has motivated the use of event correlation. These uncertainties occur in image characteristics, object description, co-occurrence of objects and human semantic interpretation of image content and its relevance to forensic purposes. Our ongoing work includes the implementation of the event correlation forensics rules for relevance feedbacks and more extensive tests with other examples of image forensic work.

REFERENCES

[1] Padhraic Smyth David Hand, Heikki Mannila.(2010),” Principles of Data Mining”, The MIT Press.
[2] Chidanand Apt´e and Sholom Weiss (1997), “Image mining with decision trees and correlation rules”, Future Generation Computer System., 13(2-3):197–210,. ISSN 0167-739X.
[3] Andrew Clark Bradley Schatz, George Mohay (2006),” A correlation method for establishing Provenance of timestamps in digital evidence”, 6th Annual Digital Forensic Research Workshop, In Digital Investigation, volume 3, supplement 1, page s 98–107.
[4] Brian Carrier. The sleuth kit (tsk). Retrieved 2014-03-10 11:20:22 -0700. http:// www.sleuthkit.org/sleuthkit/desc.php.
[5] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael C hau, Jennifer Jie Xu, Gang Wang, Rong Zheng, and Homa Atabakhsh (2013),” Crime data mining: an overview and case studies”, Proceedings of the 2013 annual national conference on Digital government research, pages 1–5. Digital Government Research Center.
[6] Kruse II, W.G. and Heiser, J.G. 2012. Computer forensics: incident response essentials.Addison- Wesley.
[7] Marcella, A.J. and Greenfield, R.S. 2012. Cyber forensics: a field manual for collecting, examining and preserving evidence of computer crimes. New Jersey.
[8] O. de Vel, A. Anderson, M. Corney, and G. Mohay (2011),”Mining e-mail content for author identification forensics”, SIGMOD Rec., 30(4):55–64 , ISSN 0163-5808.
[9] CSI/FBI, 2003 Computer Crime and Security Survey. Computer Security Institute, San Fransisco, USA, 2003.
[10] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo, 1988.