Image Cataloguing Tool using Descriptor for Forensic Application

Shubhangi T Patil
Department of Computer Science & Engineering
G. H. Raisoni Collge Of Engineering
Nagour, India

Nikita A Chavan
Department of Computer Science & Engineering
G. H. Raisoni Collge Of Engineering
Nagpur, India

ABSTRACT

Child pornography is the fastest growing criminal activity in all over the world. This crime is regarded as being extremely harmful; its prosecution is of the highest priority for police forces and law enforcement organizations around the world. The law of possession and distribution of CP images not only applies to regular citizens, but also to police units that they not allowed to gather and catalogue evidence for future reference. To overcome from this problem, the paper presents a tool which is able to catalogue high- and low-level metadata of the evidence material and provides fast search to retrieve such evidence in suspect's system. It also provides the relation between the previous cases and current cases. That would be helpful to provide the protection to victims, capturing the offenders in effective manner and also reduce the tedious works of police units.

Keywords—Child pornography; Forensic analysis; MPEG-7 descriptor

1. INTRODUCTION

The natural evolution of computers and the popularization of the Internet allowed a higher speed of communication between people. Now a days, the exchange of information between people all over the world is vast, mainly through the World Wide Web (WWW), the peer-to-peer networks, e-mail, and instant messaging services. This facility has also been used in the propagation of illegal digital multimedia content, especially images and videos, involving child pornography. Until November 25th, 2008, Brazilian law only criminalized the disclosure of such material, under Article 241 of the *Child and Adolescent Law*. After that date, with the publication of *Law 11,829*, possession of files of this nature also became a crime, with the inclusion of Article 241-B. The law applies not only to regular citizens, but also to police units. Thus, requests for forensic analysis in computer storage media such as hard disk drives, pen drives, optical media, and memory cards in order to verify the presence and distribution of child pornography are becoming more frequent. A recent forensic study showed that 148 (one hundred and forty-eight) files containing sexual abuse of a teenage girl were found after analysis of more than 300,000 image files and 1,100 video files. This shows the great difficulty of finding such files, because the analysis of images and videos is usually done visually.

Local regulations force police officers to destroy all evidence after the investigation is completed. This significantly complicates the gathering and presenting of evidence in police investigations. Due to this it is not possible for police forces to make connection between individual cases in order to find evidence in suspect's system.

To overcome from this problem, the paper presents a tool which is able to catalogue the image of child pornography and also it can be able search the images related to child pornography in suspected system. In the proposed system, police officer can archive the metadata of media obtained during investigation and store it in database using INDEXER component. When they have access to the suspect's system, they can use SEARCHER component to find all the media that is identical or visually similar to the samples present in the database.

This tool is classified into a set of two applications utilizing the Query by Example (QbE) approach. Several such applications have already been demonstrated, including a well-known, generic MIRROR content-based image retrieval system [4], it as well as more specific systems such as GAMA, designed for accessing libraries of media art [12]. The concept of using hash sets to identify suspicious files in digital forensics has been in use for a number of years, and it is built into numerous forensic tools such as EnCase and FTK (Forensic Tool Kit).

The above mentioned tools only allow the retrieval of images that are identical (they utilize MD5 sums), and also allows the retrieval of similar images, since it utilizes MPEG-7 descriptor values [15]. Although both the forensic hash tools and the MPEG- 7 standard are well-known techniques, their combination is novel.

2. RELATED WORK

Child pornography is unlike most crimes local police departments handle. Local citizens may access child pornography images that were produced and/or stored in another city or on another continent. Alternatively, they may produce or distribute images that are downloaded by people thousands of miles away. An investigation that begins in one police district will almost certainly cross jurisdictional boundaries. The following approaches are used to combat with child pornography.

A. *Host-based approaches*

One area related to child pornography detection is *content-based image and video retrieval* (CBIR). This area has become more important considering the increased use and sharing of digital visual (image and video) files. The associated complexity of the problem is to be able to retrieve visual files based on their semantics. The semantics of a file are determined according to a set of characteristics (e.g., colour contrast, shapes, etc.) learned a-priori from similar files.

The CBIR approach does not only need to focus on recognizing child pornography itself, but can be used to match pictures taken within the same environment in order to identify the origin of the material [16]. CBIR could potentially be combined with the concept of "white worms" that scan the Internet in search of child pornography, and report suspicious content to law enforcement. However, this technique has strong legal implications and has proven itself controversial in the past.

Numerous software packages are available on the market, and it is becoming increasingly problematic to determine which filter is robust enough to detect the highest incidence of child pornography on the Internet. CyberPatrol, Net Nanny, SurfWatch and CYBERsitter are some of the popular ones available, and usually come with a pre-determined blacklist that restricts access to content that is known to be of questionable nature. These lists can be customized to suit the child as well as the preferences of the parents, but need to be continually updated.

Another approach is to allow the ISP to provide this service using their already-customized list. This requires no installation on the desktop, which may be the preferred option for some parents who are not technically inclined. The other method is to list appropriate content on a white-list, allowing restricted access to that information while blocking access to all other content. Site labels with ratings are assigned to website content, and these can be used to determine access rights. Automated scanning of text, whether on a website or as part of a search query, is another technology employed to determine illicit content. A server log file can be used to store all activity records that were collected from browsing activities online. Another method is to implement passwords, encryption techniques and other authorization controls such as credit card numbers to access certain services or sites.

Host-based solutions suffer limitations. Not all hosts are accessible or open to scanning, nor do all individuals know how to enable parental controls or other safeguards. A network infrastructure approach could overcome many of these issues.

B. Network-based approaches

There have been a number of important efforts to classify packets at the network level. There are intrusion detection systems, which perform packet reassembly; however, the imposed overhead impedes the viability of these techniques in actual use within core routers. Upon the impracticability of packet reassembly, it has been shown that statistical analysis can effectively classify packets.

This scheme is based on the Naïve Baye classification technique, and is capable of classifying packets into a number of application categories, including P2P, e-mail, and multimedia files. Moore *et al*. [21] presents a packet classification scheme, which incorporates a number of classification methods into a single system that combines them according to certain rules. Although these two classification strategies use higher order Markovian models, they are prone to ignore data distributions that quickly change in time, e.g. data showing some degree of non-stationarity. Furthermore, textual files can be flagged as censured material using e-mail surveillance techniques.

It is important to realize that information is passed along the Information Highway from host to destination not as entire pieces of data but in information packets that are disassembled and reassembled. When the data to be transferred over the Net exceed a predefined maximum packet size, the data are fragmented into smaller Information Protocol (IP) packets, which are reassembled at the destination end. Routers do not perform packet reassembly; therefore, intermediate routers participating in a visual file transfer may not see the whole visual file within a single packet. Regarding performance constraints, visual files classification is time-consuming and routers may not be able to cope with high volumes of traffic.

C. Censured peer-to-peer traffic

A number of P2P program vendors have entered alliances in order to combat the issue of child pornography [17]. Most programs used for file sharing are not in any way part of such an alliance or any others. The state of P2P programs is rapidly changing to include heavy encryption and anonymity features. One of the largest of these networks is facilitated by a program called “Share”, which is one of the most popular P2P programs in use in Japan.

The P2P field enjoys much attention within research circles, and it is conceivable that it may become a primary distribution point for child pornography in the future. Features such as censorship-resistant publishing make it possible to anonymously and systematically distribute sets of files on networks – an environment that can be abused for malicious purposes [18]. The problem presented by the P2P field has different possible approaches, since the various P2P protocols in existence have different methods to upload or download material. As such, targeting any single P2P network requires detailed knowledge of the protocol in use. Network infrastructure techniques can be developed to analyze traffic without knowledge of overlying protocols or the specifics of networks such as the ones created by P2P applications. Any solution targeted at a specific protocol, application, or network will have limited applicability, and likely decreasing utility as offenders move to new ways of distributing child pornography.

It is important to consider the case of anonymous and encrypted file P2P applications such as Winny [19]. These applications introduce additional complexity for detection and identification. Encrypted payloads are not susceptible to CBIR or e-mail surveillance techniques. Ohzahata *et al.* [20] present identification mechanisms based on TCP handshake behavior. Their system resides inside an autonomous system, and is capable of detecting encrypted P2P file transfers. Attribution of anonymous P2P nodes at the network infrastructure level is challenging, and is a problem yet to be addressed. Identifying encrypted traffic is out of the scope of this technique, and is an area that we plan to explore in the future.

All the above techniques provides searching of illegal images from suspected system but not provide any database so that they can help to police unit for the further investigation or provide the protection to victims for capturing the offenders in effective manner.

3. CONCEPT OF APPLICATION

In this application INDEXER is designed to be used at police head quarter. The police have at their disposal sets of images containing child pornography from various sources; include ongoing investigations and international channels of cooperation. These sets are given as input to INDEXER. The INDEXER processes the images and converts that set of images in the form of image descriptor. The descriptor set consists of MD5 hashes and MPEG-7 descriptors. This process is one way process which means that the images cannot be recreated either from the hashes or from the descriptors. Those descriptors are finally store in hash database which is then used by SEARCHER application.

The database is kept centralized. The SEARCHER application is used for searching the suspect’s system. It uses the database of hashes and descriptors created by the INDEXER. All images that are identical (utilisingMD5 sums) or similar (utilisingMPEG-7descriptor values) will be retrieved and presented alongside the database information to the officer performing the search. This allows the police to draw conclusions regarding the possible sources of the images and their distribution paths.

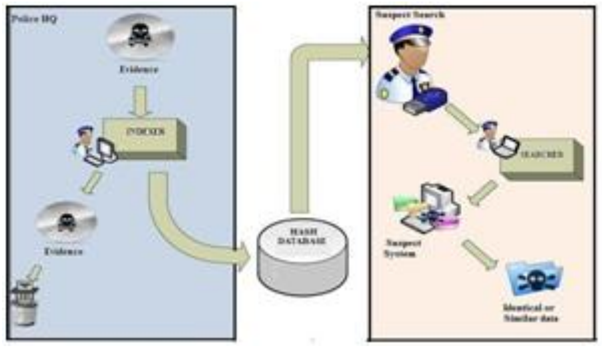


Fig.1 System Architecture

The database is kept centralized. The SEARCHER application is used for searching the suspect's system. It uses the database of hashes and descriptors created by the INDEXER. All images that are identical (utilisingMD5 sums) or similar (utilisingMPEG-7descriptor values) will be retrieved and presented alongside the database information to the officer performing the search. This allows the police to draw conclusions regarding the possible sources of the images and their distribution paths

4. OUR APPROACH

As defined above the first step for our application is to create the INDEXER. It consists of MPEG-7 descriptors and MD-5 hashes.

A. Descriptor Creation

The Moving Picture Experts Group (MPEG) has defined several visual descriptors in their standard referred to as MPEG-7 standard. MPEG-7 Visual Standard specifies a set of descriptors that can be used to retrieve similar images from digital photo repository. The descriptors that were proposed by MPEG-7, for indexing and retrieval, maintain a balance between the size of the feature and the quality of the results. These descriptors appear to be able to describe satisfactorily the visual content of the image. The MPEG-7 standard has tested the most efficient procedure to describe the color and has selected those that have provided more satisfactory results. A digital image can be described as a point matrix where each point (*pixel*) represents a color.

The RGB color space was originated from the old CRT (Cathodic Ray Tube). It defines the color representation by combining three primary colors: red (R), green (G), and blue (B). The scale of each of these components varies from 0 to 255. This is the most common color space used for storing digital image representation. In RGB, brightness and color are coupled and thus not suitable for color segmentation in images with unknown light conditions. To reduce the effect of luminance on the color representation, it is possible to use the normalized RGB, which consists of a transformation from RGB through a normalization process.

The SEARCHER application is based on the filtering process. This process is useful for separating human pictures and child pictures from the system.

B. Face detection and Face Recognition

As face detection is the first step of any face processing system, it finds numerous applications in face recognition, face tracking, facial expression recognition, facial feature extraction, gender classification, clustering, attentive user interfaces, digital cosmetics, biometric systems, to name a few. Face recognition has received much attention due to its potential values for applications as well as theoretical challenges. However, despite the research advances over these years, face recognition is still a highly difficult task in practice due to the large variability of the facial images. The variations between images of a same face can generally be divided into two categories: *the appearance variations*, and *the man-made variations*. The appearance variations include facial expression, pose, aging, illumination changes, etc. And the man-made variations are mainly due to the imperfections of the capture devices and image processing technologies, e.g. the noises from the cameras and the face registration error resulting from imperfect face detections.

Most detection systems carry out the task by extracting certain properties (e.g., local features or holistic intensity patterns) of a set of training images acquired at a fixed pose. To reduce the effects of illumination change, these images are processed with histogram equalization or standardization (i.e., zero mean unit variance). Based on the extracted properties, these systems typically scan through the entire image at every possible location and scale in order to locate faces. The extracted properties can be either manually coded (with human knowledge) or learned from a set of data as adopted in the recent systems that have demonstrated impressive results.

Many algorithms implement the face-detection task as a binary pattern-classification task. That is, the content of a given part of an image is transformed into features, after which a classifier trained on example faces decides whether that particular region of the image is a face, or not. Often, a window-sliding technique is employed. That is, the classifier is used to classify the (usually square or rectangular) portions of an image, at all locations and scales, as either faces or non-faces (background pattern).Images with a plain or a static background are easy to process.

C. Nudity Detection

Nudity detection is recognized as an important step towards limiting the proliferation of unwanted and damaging information. Nudity Detection is used for the extraction of image feature. It is possible to compare the size of the largest skin region found with the image size. The amount of identified skin areas is another important feature. Small skin regions in large quantities are not typical of nudity images. The percentage of skin in an image is also an indicative of the presence of nudity.

A basic nudity filtering solution requires the analysis of an image to determine the presence of objects in the image that may signify nudity or inappropriate content.

D. Age Classification

The theory has only been implemented to classify input images into one of three age groups: babies, young adults, and senior adults. The computations are based on cranio-facial development theory and skin wrinkle analysis. In the implementation, primary features of the face are found first, followed by secondary feature analysis. The primary features are the eyes, nose, mouth, chin, virtual-top of he head and the sides of the face. From these features, ratios that distinguish babies from young adults and seniors are computed. In secondary feature analysis, a wrinkle geography map is used to guide the detection and measurement of wrinkles. The wrinkle index computed is sufficient to distinguish seniors from young adults and babies.

E. Rule Based Criteria

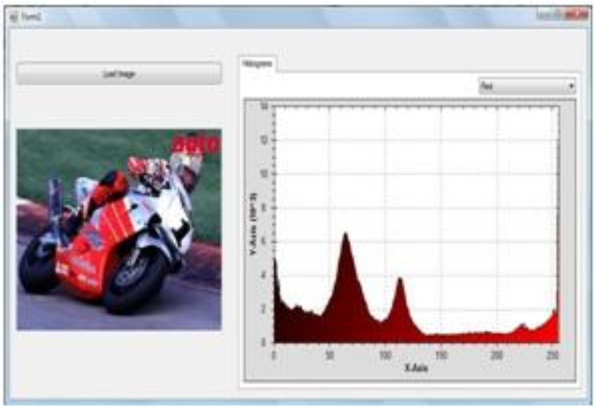
Rule based systems. Such systems deal with decision making taking into consideration a description given by the user, and data and rules, provided by the system. Most of these systems are informational-instructive with (or without) an additional diagnosis component.

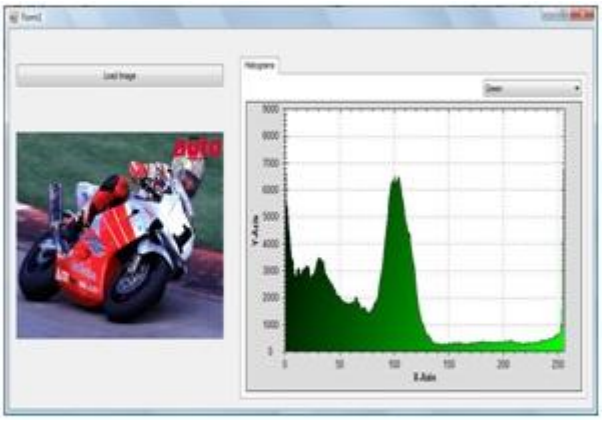
5. EXPERIMENTAL WORKS AND RESULTS

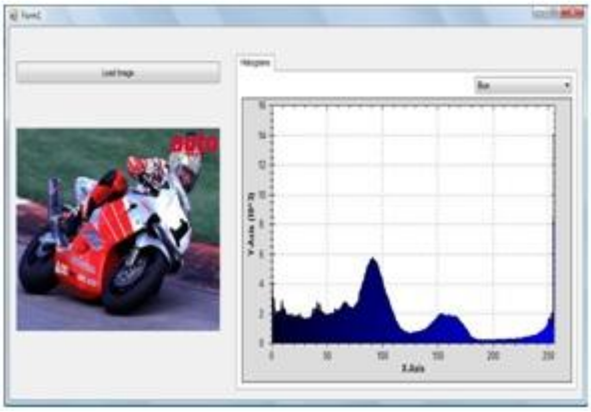
Note that we have not detected the complete image of child pornography. In our experimental work we have done only up to the face recognition process. That means we have firstly create the descriptor for the image in INDEXER component by using MPEG-7 one way descriptor. In SEARCHER component we performed face recognition process on input image to determine whether the input image is of child or adult.

A. Descriptor Creation

Descriptors are the first step to find out the connection between pixels contained in a digital image and what humans recall after having observed an image or a group of images after some minutes. This is one way descriptor that is after creation if descriptor we cannot create the original image from that descriptor. The descriptors of the particular image are as follows.

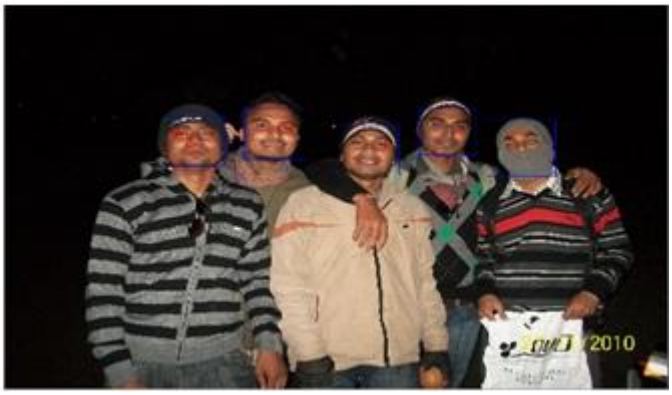






B. Face detection and Face Recognition

Face detection is the first process in the SEARCHER component. For the input image we firstly detect the face in the input image. After that on this image we perform face recognition process and next we perform nudity detection and classify the image on the basis of babies, young adults, and senior adults.



6. CONCLUSION

This paper introduced an image cataloguing tool that creates a centralized hash database of child pornography images, which would not be allowed by police unit to store and the searching of those images in the suspected system in fast way. It also provides the relation between the previous cases and current cases. That would be helpful to provide the protection to victims, capturing the offenders in effective manner and also reduce the tedious works of police units. Implementation of this system is based upon the INDEXER and SEARCHER component that makes possible for the police unit to set up database that contain the images of child pornography. The concept of storing information on child pornography is not new. The basis for identifying images with child abuse material is a hash value calculated using a dedicated file. It is the simplest yet not very progressive method of investing such images. The main advantage of the system is enabling police forces to search for files resembling each other in their content based upon MPEG-7 descriptor. This functionality should allow the police to expand their investigation and unveil more evidence.

7. REFERENCES

[1] Nadernejad E. Sharifazadesh S. Hassanpur H (2008) Edge detection techniques: evalution and compression. Appl Math Sci 2(2-31):1507-1520

[2] Rogers MK, Goldman J, Mislam R, Wedge T, Debrota s (2006) Computer forensic field triage process model In: Conference on digital forensics, security and law.

[3] Wong KM, Cheung KW, Po LM (2005) Mirror: and interactive content based image retrieval system In: ISCAS (2) IEEE, pp 1541-1544.

[4] Pham DT, Ghanbarzadeh A(2007) Multiobjective optimization using the Bees algorithm. In: Innovative production machines and systems virtual conference.

[5] Pham DT, Ghanbarzadeh A, Koc, E, Otri S, Rahim S, Zaidi M (2006) The Bees Algorithm—a novel tool for complex optimizationproblem. In: Proceeding of the 2nd international virtual conference on intelligent production machines and system (IPROMS 2006). Elsevier, Oxford, pp 454-459.

[6] AccessData: Forensic Toolkit (FTK) computer forensic software http:\\accessdata.com/products/forensicinvestigation/ftk. Accessed 30 Sept 2011.

[7] Mohand-SaoEd Hacid. Cyril Decleir, and Jacques Kouloumdjian, A Databse Approach for Modeling and Querying Video Data IEEE Transaction on Knowledge And Data Engineering, Vol. 12, No. 5, September/October 2000 729.

[8] Eleuterio P, Polastro M (2010) Optimization of automatic nudity detection in high-resolution images with the use of NuDetective Forensic Tool. In: Proceeding of the fifth international conference on forensic computer science—ICoFCS 2010, Brasilia, Brazil.

[9] Osama A. Lotfallah, Martin Reisslein, *and Sethuraman Panchanathan, Fellow,* Adaptive Video Transmission Schemes Using MPEG-7 Motion Intensity Descriptor IEEE Transactions On Circuits And Systems For Video Technology, Vol. 16, No. 8, August 2006 929.

[10] International Center for Missing and Exploited Children (2006) Child pornography: model legislation and global review, 6th edn. http://books.google.pl/books?id=8iqqOwAACAAJ. Accessed 30 Sept2011

[11] Stefan Berchtold, Daniel A. Keim, Indexing the Solution Space: A New Technique for Nearest Neighbor Search in High-Dimensional Space IEEE Transactions On Knowledge And Data Engineering, Vol. 12, No. 1, January/February 2000 45

[12] Ludtke A, Gottfried B, Herzog O, Ioannidis G, Leszczuk M, Simko V (2009) Accessing libraries of media art through metadata. In: International workshop on database and expert systems applications, pp 269273

[13] Javier Ruiz Hidalgo, *and Philippe Salembier,* On the Use of Indexing Metadata to Improve the Efficiency of Video Compression IEEE Transactions On Circuits And Systems For Video Technology, Vol. 16, No. 3, March 2006

[14] Pardo A (2006) Probabilistic shot boundary detection using interframe histogram differences. In: CIARP'06: Proceedings of the 11th Iberoamerican conference on progress in pattern recognition, image analysis and applications. Springer-Verlag, Berlin, Heidelberg, pp 726–732. doi:10.1007/11892755_75

[15] Manjunath BS, Salembier P, Sikora Th (2002) Introduction to MPEG-7 media content description interface. Wiley, Chichester

[16] Tanase-Avatavului, M., 2005. Shape Decomposition and Retrieval. *Doctoral Thesis, University of Utrecht*. Available online: http://www.cs.uu.nl/groups/AA/multimedia/publications/pdf/mindshade.pdf [Accessed: June 24, 2005]

[17] Borland, J., 2004. P2P Group Launches Site to Combat Child Porn. *CNet News.com*. Available online: http://news.com.com/P2P+group+launches+site+to+combat+child+porn/2100-1025_3-5488290.html [Accessed: June 24, 2005]

[18] Tand, C., et al., 2003. Peer-to-Peer Information Retrieval Using Self Organizing Semantic Overlay Networks.*Proceedings of the ACM SIGCOMM 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, Karlsruhe, Germany, pp. 175-186.

[19] Wikipedia contributors, 2005. Winny. *Wikipedia, The Free Encyclopedia*. Available online: http://en.wikipedia.org/w/index.php?title=Winny&oldid=31980964 [Accessed: June 24, 2005]

[20] Ohzahata, S. et al., 2005. A Traffic Identification Method and Evaluations for a Pure P2P Application. *Proceedings of Passive & Active Networks Measurement Workshop*, Boston, USA, pp. 55-68.

[21] Moore, A.W. and Papagiannaki, K., 2005. Toward the accurate Identification of network applications," *Proc. Passive & Active Networks Measurement Workshop*, Boston, USA, 2005